Semi-parametric copula sample selection models for count responses*

Giampiero Marra Karol Wyszynski

May 8, 2016

Abstract

In observational studies, a response of interest (as well as some individual level characteristics) may be observed for a non-randomly selected sample of the population. In this situation, standard models such as linear and probit regressions will yield biased and inconsistent parameter estimates. Selection models can address this issue and mainly consist of two regressions: a binary selection equation which determines whether the statistical units will enter the sample, and an outcome equation which models the response. While sample selection models for continuous and binary outcomes have been widely studied in the literature, the case of count response has not received as much attention. Sample selection models for count data which allow for the use of potentially any discrete distribution, non-Gaussian dependencies between the selection and outcome equations, and flexible covariate effects are introduced. The estimation algorithm is based on the penalized likelihood estimation framework. The method is illustrated in simulation and using data from a United States Veterans Administration Survey.

Keywords: Non-random sample selection; copula; penalized regression spline; count response.

1 Introduction

Non-random sample selection arises when individuals select themselves into (out of) the sample based on features that are observed and unobserved. In this case, statistical analysis based on commonly known models such as linear and probit regressions will yield biased and inconsistent parameter estimates. One way of addressing this issue is to employ sample selection models.

The motivating example of this work stems from data collected through the 2001 United States Veterans Administration Survey (USVA, 2001). Here, the interest is in estimating the impact of certain observed patients' characteristics on number of visits in Veterans Administration (VA) and non-VA medical facilities and the predicted average number of visits (Lahiri & Xing, 2004). This study is challenging because of the likely presence of relevant unobserved factors (e.g., attitude towards health related risks); neglecting the difference in the unobserved attributes of the individuals who used the facilities and those who did not use them may have an adverse effect on parameter

*Authors' e-mails: giampiero.marra@ucl.ac.uk, k.m.wyszynski@gmail.com. Address: Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

estimation (Lahiri & Xing, 2004; Trivedi & Zimmer, 2007, p. 76). In this case, an appropriate approach such as the sample selection methodology is required to obtain consistent estimates.

Sample selection models, also known as Heckman-type models, were first introduced by Gronau (1974), Lewis (1974) and Heckman (1976), and discussed more thoroughly in Heckman (1979). They typically consist of a selection equation and an outcome equation. The former models whether an observation will be missing on the response (e.g., decision to use the facilities) and is usually achieved using a probit regression. The latter models the response of interest (e.g., number of hospital visits) and the type of regression employed typically depends on the nature of the response. The two equations are allowed to be associated which will be crucial when non-random sample selection is based on unobservables. The literature on models tackling selection bias is vast and without claim of exhaustiveness we mention below some interesting variants. Chib et al. (2009) and Wiesenfarth & Kneib (2010) introduced Bayesian sample selection models which allow researchers to flexibly estimate covariate effects, whereas a frequentist counterpart was introduced by Marra & Radice (2013a). Li (2011) considered the situation in which there is more than one selection mechanism, and Omori and Miyawaki (2010) extended selection models to allow threshold values to depend on individuals' characteristics. These models have also been compared to principal stratification in the context of causal inference with non-ignorable missingness (Mealli and Pacini, 2008). Liu et al. (2012) employed sample selection models based on a three equation system, whereas Marra & Radice (2013b) focused on binary outcomes. Greene (1997), Terza (1998) and Miranda & Rabe-Hesketh (2006) discussed the case of count responses; the approaches of these authors have the main drawbacks that (typically computationally expensive) quadrature or simulation methods have to be employed to obtain certain key quantities required for model fitting, data-driven semi-parametric effects are not allowed for and, generally speaking, the unconditional distribution of the response of the outcome equation is unknown. Taking a different view to the problem of non-observable response, it is possible to assign a zero value to the outcome whenever an observation on it was not generated (e.g., the individual did not use the facilities) and assume that such a value is "genuine". In this case, two-part and zero inflated models may, for example, be employed (see Humphreys (2013), Lambert(1992) and references therein). The former consists of fitting two regressions, one for modeling the occurrence of zeros and the other for modeling non-zero values. The latter uses a mixture of Bernoulli and discrete distributions which accounts for excess of zeros. These approaches essentially aim at modeling zero and non-zero values, all of which observed, instead of dealing with missing observations on the response which is when selection models are most useful. Therefore, the key question for the researcher is to determine why zero values are present in the response and then choose an appropriate methodology.

Selection approaches that rely on the commonly used bivariate normality assumption are often criticized since if this fails to hold then the resulting estimator will not yield consistent estimates (e.g., Pigini, 2012; Smith, 2003). The literature offers some alternatives to the assumption of Gaussianity, including non/semi-parametric and copula approaches. The former tend to be computer-intensive and typically do not allow for much flexibility in the model specification (compare: Pigini, 2012).

Furthermore, convergence problems may arise when fitting models with several types of covariate effects (Wojtyś, in press). The copula approach is more feasible as it uses maximum likelihood techniques. It also allows for simultaneous estimation of all model parameters which may lead to important efficiency gains (Smith, 2003). However, it may deliver estimators that are not consistent when the distributional assumption is not correct. Nevertheless, copulae allow for a piece-wise model specification and for many modeling options; for instance, it is possible to use any two marginal distributions when the copula linking them is Joe or Clayton. This is advantageous as the user can assess the sensitivity of results to different modeling assumptions. Genius and Strazzera (2008) pointed out that the copula approach allows for direct estimation of the dependence between the two equations, while non/semi-parametric methods do not. Hasebe & Vijverberg (2012) established a sample selection model based on copulae and the Generalized Tukey Lambda (GTL) marginal distribution. The authors argue that GTL is an appealing choice as it allows for skewness and thin and heavy tailed response behavior. Marchenko & Genton (2012) developed the selection-t model in which the errors are modeled using a bivariate t-distribution. Finally, it is worth mentioning the works of Cameron & Trivedi (2005), Cameron & Trivedi (2013) and Cameron et. al (2004) who exploited copulae for modeling count data in various contexts including non-random selectivity.

The practical relevance of the selection approach is supported by the number of hits received by Google Scholar when typing "sample selection model" (over 1200 hits since 2014, the majority of which relate to applied articles and reports). This paper contributes to the literature by introducing a flexible copula-based sample selection modeling approach for count data. The proposed method allows for the use of several (copula) dependence structures and (potentially) any discrete outcome margin (as long as the probability mass function (pmf) and cumulative distribution function (cdf) are known). Covariate effects are flexibly determined by the data using, for instance, penalized thin plate regression splines or P-splines commonly known in the context of Generalized Additive Models (e.g., Wood 2006). The proposed framework is termed as semi-parametric (following the typical convention adopted in the statistical modeling literature when covariate effects are estimated flexibly) and is not affected by the aforementioned drawbacks of available selection approaches for count data. Previous works on selection models have considered separately the use of copulae, semi-parametric covariate effects and discrete distributions. This paper brings together these strands of research which required a considerable methodological effort and some careful structuring when implementing the proposed class of models. Moreover, our approach allows one to model distribution specific parameters as functions of semi-parametric effects as advocated by Stasinopoulos & Rigby (2005) in the context of univariate generalized additive models. To the best of our knowledge, this development has never been considered in the context of the models introduced in this paper.

The remainder of this paper is organized as follows. Section 2 introduces the models and likelihood. Section 3 discusses parameter estimation which is based on penalized maximum likelihood, whereas Section 4 briefly mentions how to construct confidence intervals and carry out model selection. Section 5 presents some simulation results, and Section 6 illustrates the framework using USVA data. Concluding remarks are given in Section 7. All developments are implemented in the

`SemiParSampleSel` package (Marra et al., 2016) for the `R` environment (R Development Core Team, 2016).

2 Sample selection models for count responses

2.1 Selection and outcome equations

Non-random sample selection occurs when some observations for the response of interest, Y_{2i}, are missing not at random. In a sample selection modeling context, it is typically assumed that there is a variable Y_{1i} which governs the selection process. Using the latent variable representation, this can be written as

$$Y_{1i}^{*} = \eta_{1i} + \epsilon_{1i} = \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{z}_i + \sum_{k_1=1}^{K_1} s_{1k_1}(u_{1k_1i}) + \epsilon_{1i},\ i = 1, \ldots, n,$$

where n is the sample size, $Y_{1i}^{*} \sim N(\eta_{1i}, 1)$, $Y_{1i} = 1$ if $Y_{1i}^{*} > 0$ and 0 otherwise, η_{1i} is a linear predictor with the obvious definition, $\boldsymbol{\gamma}$ is a P_1 dimensional coefficient vector of all parametric components, $\boldsymbol{z}_i$ is a vector of (binary and categorial) covariates and the $s_{1k_1}(u_{1k_1i})$ are unknown smooth functions of the K_1 continuous covariates u_{1k_1i}. The outcome equation can be written as

$$Y_{2i} \ \sim \ \mathcal{D}(\mu_i, \sigma, \nu),$$

where

$$\mu_i = E(Y_{2i}) = \exp(\eta_{2i}) = \exp\left(\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{x}_i + \sum_{k_2=1}^{K_2} s_{2k_2}(u_{2k_2i})\right),$$

$\mathcal{D}$ is a discrete distribution (several choices will be discussed in Section 2.3), η_{2i} is a linear predictor, and σ and ν are scale and shape distribution specific parameters. The number of parameters that characterize $\mathcal{D}$ depends on the chosen distribution. Without loss of generality, we parametrize all the distributions discussed in this article in terms of μ_i, σ and ν. Vector $\boldsymbol{\beta}$ has length P_2 and represents the parameters of all parametric components, $\boldsymbol{x}_i$ is a vector of factor variables and the $s_{2k_2}(u_{2k_2i})$ are unknown smooth functions of the K_2 continuous covariates u_{2k_2i}.

The vk_v smooth component $s_{vk_v}(u_{vk_vi})$ is approximated using regression splines (Wood, 2006) as

$$\sum_{j=1}^{J_{vk_v}} \alpha_{vk_vj} b_{vk_vj}(u_{vk_vi}) = \mathbf{B}(u_{vk_vi})^{\mathsf{T}}\boldsymbol{\alpha}_{vk_v},$$

where the $b_{vk_vj}(u_{vk_vi})$ are known spline basis function, the α_{vk_vj} are regression parameters, J_{vk_v} is the number of bases used to represent the smooth term, $\mathbf{B}(u_{vk_vi}) = [b_1(u_{vk_vi}), b_2(u_{vk_vi}), ..., b_{J_{vk_v}}(u_{vk_vi})]^{\mathsf{T}}$ is a vector containing J_{vk_v} basis functions and $\boldsymbol{\alpha}_{vk_v}$ is the corresponding parameter vector. Evaluating the basis functions for each observation yields J_{vk_v} curves which multiplied by the respective coefficients and then summed will yield an estimate of $s_{vk_v}(u_{vk_vi})$ (e.g., Ruppert et al., 2003). Basis

functions should be chosen to have convenient mathematical forms and numerical properties. Our default choice is the low rank thin plate regression spline (Wood, 2006), but other options are available such as cubic regression splines (see Supplementary Material 1 for a review of splines). Each smooth function in the model is subject to the centering identifiability constraint $\sum_i s_{vk_v}(u_{vk_vi}) = 0$ (Wood, 2006).

Our implementation also allows for the specification of linear predictors for σ and ν, in the spirit of Stasinopoulos & Rigby (2005). For instance, $\sigma_i = \exp(\eta_{3i}) = \exp\left(\boldsymbol{\omega}^\mathsf{T}\boldsymbol{w}_i + \sum_{k_3=1}^{K_3} s_{3k_3}(u_{3k_3i})\right)$ and $\nu_i = \eta_{4i} = \boldsymbol{\psi}^\mathsf{T}\boldsymbol{q}_i + \sum_{k_4=1}^{K_4} s_{4k_4}(u_{4k_4i})$. Quantities $\boldsymbol{\omega}$ and $\boldsymbol{\psi}$ are parameter vectors associated with $\boldsymbol{w}_i$ and $\boldsymbol{q}_i$ (the vectors of factor variables) and the $s_{3k_3}(u_{3k_3i})$ and $s_{4k_4}(u_{4k_4i})$ are unknown smooth functions. In this case the interpretation of regressor effects may become more involved, hence such specifications need to be well justified. To avoid clutter in the notation, in the subsequent sections we will drop the subscripts referring to the number of observations.

2.2 Linking the model equations

In a sample selection model it is assumed that the selection and outcome equations are linked through unobservables; in the standard case, this link is formalized using a bivariate normal distribution. The association between a generic pair of variables (Y_1, Y_2) can be represented using a copula function (Sklar, 1959). Specifically, let $F_1(y_1)$ and $F_2(y_2)$ denote the cdfs of y_1 and y_2. Then, for a two-place copula function C, the joint cdf has the representation

$$F(y_1, y_2) \quad = \quad C(F_1(y_1), F_2(y_2); \theta),$$

where θ is a dependence parameter. When F_1 or F_2, or both cdfs are discrete, the copula is unique only on the closure $\text{Ran}(F_1)\times\text{Ran}(F_2)$, where $\text{Ran}(\cdot)$ denotes the range of its argument (e.g., Nikoloulopoulos & Karlis, 2010). Also, for discrete margins, if one knows, for instance, F_2 then the "corresponding F_2^{-1}" would be the integer among the possible F_2^{-1} values. However, the lack of uniqueness for discrete-discrete or continuous-discrete margins is a theoretical aspect that needs to be confronted in analytical proofs and does not affect empirical applications; practitioners are more interested in choosing the appropriate copula and its margins rather than knowing the exact mathematical form of the joint distribution which is a requirement for finding a unique copula (Trivedi and Zimmer, 2007). Since in some cases θ may not have a straightforward interpretation, θ can be converted into Kendall's τ which lies in the range $[-1, 1]$ and has a universal interpretation for all copulae. As pointed out by Genest and Neslehova (2007), for continuous and discrete margins Kendall's τ is not a margin-free measure of dependence and should therefore be used with caution especially when the choice of margins is based on dependence considerations. However, this is less of a concern here as in our approach copulae represent a means to link the selection and outcome equations where the marginal cdfs are chosen by looking at the responses of interest. One of the main advantages of copulae is that they allow for a piece-wise model specification; in our context, this means that it is possible, for instance, to use normal-Poisson margins when the copula is Gaussian, Joe or Clay-

ton. The copulae currently available in `SemiParSampleSel` and the respective conversions from θ to Kendall's τ are summarized in Table 1.

Name	**Copula** $C(u, v; \theta)$	**Parameter space of** θ	**Parameter space of Kendall's** τ	**Kendall's** τ **in terms of** θ
FGM	$uv(1+\theta(1-u)(1-v))$	$-1 \leq \theta \leq 1$	$-2/9 \leq \tau \leq 2/9$	$\frac{2}{9}\theta$
Normal	$\Phi_2(\Phi^{-1}(u), \Phi^{-1}(v); \theta)$	$-1 \leq \theta \leq 1$	$-1 \leq \tau \leq 1$	$\frac{2}{\pi}\arcsin(\theta)$
AMH	$uv/(1-\theta(1-u)(1-v))$	$-1 \leq \theta \leq 1$	$-0.1817 \leq \tau < \frac{1}{3}$	$1-\frac{2}{3\theta^2}(\theta+(1-\theta)^2 \log(1-\theta))$
Clayton	$(u^{-\theta}+v^{-\theta}-1)^{-1/\theta}$	$0 < \theta < \infty$	$0 < \tau < 1$	$\frac{\theta}{\theta+2}$
Frank	$-\theta^{-1}\log(1+(e^{-\theta u}-1)$ $(e^{-\theta v}-1)/(e^{-\theta}-1))$	$\theta \in \mathbb{R} \setminus \{0\}$	$-1 < \tau < 1$	$1-\frac{4}{\theta}[1-D_1(\theta)]$
Gumbel	$\exp(-((-\log u)^\theta+(-\log v)^\theta)^{1/\theta})$	$1 \leq \theta < \infty$	$0 \leq \tau < 1$	$1-\frac{1}{\theta}$
Joe	$1-((1-u)^\theta+(1-v)^\theta-$ $(1-u)^\theta(1-v)^\theta)^{1/\theta}$	$1 < \theta < \infty$	$0 < \tau < 1$	$1+\frac{4}{\theta^2}D_2(\theta)$

Table 1: Examples of families of bivariate copulae. u and v represent marginal cdfs. $\Phi_2(\cdot,\cdot;\theta)$ denotes bivariate standard normal cdf with correlation coefficient θ, and $\Phi^{-1}(\cdot)$ is the inverse cdf of a standard normal. $D_1(\theta) = \frac{1}{\theta}\int_0^\theta \frac{t}{\exp(t)-1}dt$ is the Debye function and $D_2(\theta) = \int_0^1 t\log(t)(1-t)^{\frac{2(1-\theta)}{\theta}}$.

The Clayton, Joe and Gumbel copulae can be rotated by 90, 180 and 270 degrees. These rotations are defined as

$$\begin{aligned} C_{90} &= v - C(1-u, v; \theta), \\ C_{180} &= u + v - 1 + C(1-u, 1-v; \theta), \\ C_{270} &= u - C(u, 1-v; \theta), \end{aligned}$$

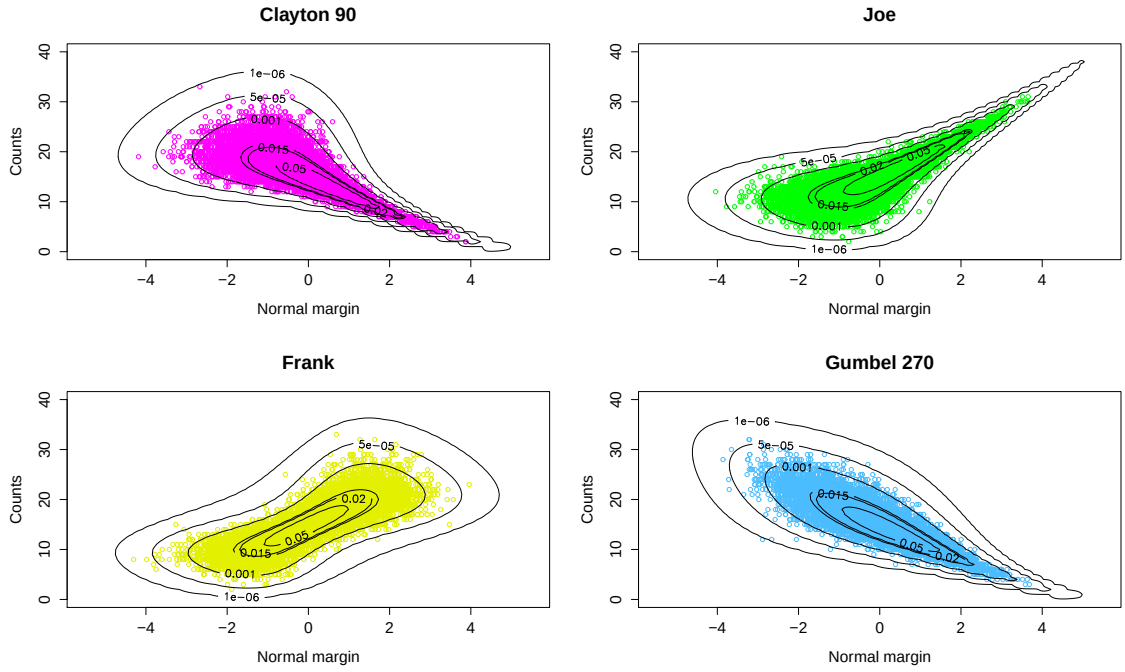
where u and v are the margins (Brechmann & Schepsmeier, 2013). This approach is useful when adopting copulae that do not have full coverage. For instance, Clayton can only capture positive dependence. If the association between the selection and outcome equations is believed to be negative, then Clayton rotated by either 90 or 270 degrees will be more appropriate. Figure 1 illustrates the Clayton 90, Joe, Frank and Gumbel 270 copulae.

It is worth mentioning that our implementation allows the copula dependence parameter to be modeled as a function of covariates (in a similar fashion as it can be done for σ and ν). In general, the choice of specification for θ should be driven by subject matter knowledge and a balance between parsimony and a reasonable reflection of the behavior of the selected observations. The linear predictor of the dependence parameter would model an unobserved selection process and therefore such an approach only makes sense from an estimation perspective if there are, for example, groups of individuals for which there is a clear rationale for expecting heterogeneous selection mechanisms.

2.3 Outcome margins

The proposed approach can incorporate a wide range of parametric discrete margins, hence allowing for greater flexibility in modeling responses. For instance, if the Poisson distribution is not suitable

Figure 1: Contour plots for Clayton 90, Joe, Frank and Gumbel 270. Bivariate densities were obtained using the principle described in Section 2.4. 20,000 deviates were generated for each copula and plotted against the contours. The selection equation margin (x-axis) is standard normal, whereas the outcome equation margin (y-axis) is Poisson with $\mu = 15$. Kendall's τ was set to 0.7 and -0.7 for the rotations by 90 and 270 degrees. The plots display different copula shapes. For instance, Joe shows a greater tail dependence in the upper right corner, whereas Clayton 90 shows a greater tail dependence in the lower right corner.



for the data at hand then the negative binomial may be chosen instead. This includes a scale parameter σ which allows for overdispersion. The outcome margin distributions adopted here are from the `gamlss R` package (Stasinopoulos & Rigby, 2007). The distributions implemented in `gamlss` are parametrized with respect to location, scale and shape which, as mentioned previously, we will denote as μ, σ and ν respectively. The parametrization for μ allows us to specify the outcome equation in the desired way since $E(Y) = \mu = \exp(\eta)$ where η is a generic linear predictor. Table 2 summarizes the distributions available in `SemiParSampleSel`. All distributions but Poisson have a scale parameter and only Delaporte and Sichel have a shape parameter with different ranges of admissible values.

Figure 2 shows some examples of pmfs for the Poisson, negative binomial, Delaporte, Poisson inverse Gaussian and Sichel distributions.

Name	μ	σ	ν	**pmf**
Poisson	$(0, \infty)$	-	-	$\frac{e^{-\mu}\mu^y}{\Gamma(y+1)}$
Negative Binomial	$(0, \infty)$	$(0, \infty)$	-	$\frac{\Gamma(y+1/\sigma)}{\Gamma(y+1)\Gamma(1/\sigma)}[\frac{(\mu\sigma)^y}{(\mu\sigma+1)}]^{y+(1/\sigma)}$
Delaporte	$(0, \infty)$	$(0, \infty)$	$(0, 1)$	$\frac{e^{-\mu\nu}}{\Gamma(1/\sigma)}[1+\mu\sigma(1-\nu)]^{-1/\sigma}S$, $S = \sum_{j=0}^{y}\binom{y}{j}\frac{\mu^y\nu^{y-j}}{y!}[\mu+\frac{1}{\sigma(1-\nu)}]^{-j}\Gamma(1/\sigma+j)$
Poisson inverse Gaussian	$(0, \infty)$	$(0, \infty)$	-	$\left(\frac{2\alpha^{1/2}}{\pi}\right)\frac{\mu^y e^{1/\sigma}K_{y-1/2}(\alpha)}{(\alpha\sigma)^y y!}$, $\alpha^2 = \frac{1}{\sigma^2}+\frac{2\mu}{\sigma}$
Sichel	$(0, \infty)$	$(0, \infty)$	$(-\infty, \infty)$	$\frac{\mu^y K_{y+\nu}(\alpha)}{c^y(\alpha\sigma)^{y+\nu}y!K_\nu(\frac{1}{\sigma})}$, $\alpha^2 = \frac{1}{\sigma^2}+\frac{2\mu}{c\sigma}$

Table 2: Summary of some discrete distributions. These are parametrized in terms of μ, σ and ν which represent location, scale and shape, respectively. For all distributions, $E(Y) = \mu = \exp(\eta)$ where η is the linear predictor. For the Poisson inverse Gaussian and Sichel distributions, $K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1}\exp(-\frac{1}{2}t(x+x^{-1}))dx$ is the modified Bessel function of the third kind. Note that $c = K_{\nu+1}(\frac{1}{\sigma})[K_\nu(\frac{1}{\sigma})]^{-1}$.

2.4 Likelihood

The likelihood of a sample selection model can be formulated generically as (Smith, 2003)

$$\begin{aligned} L &= \prod_0 \Pr(Y_1 = 0)\prod_1 P(Y_2 = y_2, Y_1 = 1) \\ &= \prod_0 \Pr(Y_1^* \leq 0)\prod_1 f_{2|1}(y_2|y_1^* > 0)\Pr(Y_1^* > 0), \end{aligned}$$

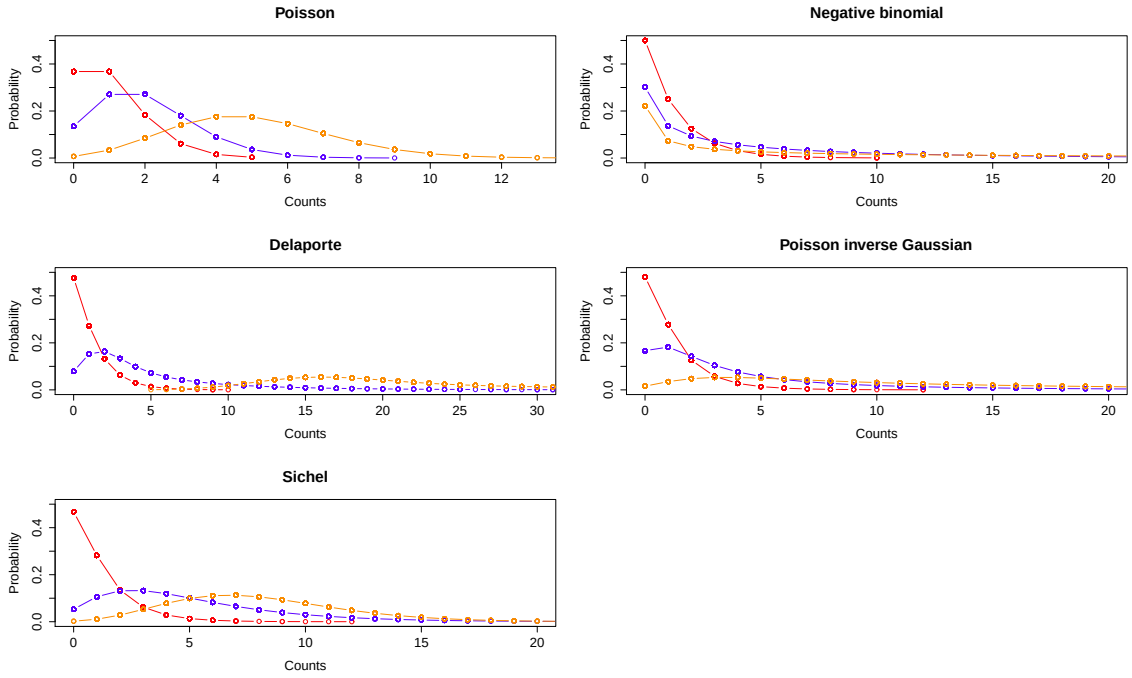
where

$$\begin{aligned} f_{2|1}(y_2|y_1^* > 0) &= \frac{\partial}{\partial y_2}\frac{F_2(y_2) - F(0, y_2)}{F_1(1)} \\ &= \frac{1}{1-F_1(0)}\frac{\partial}{\partial y_2}(F_2(y_2) - F(0, y_2)) \qquad (2.4.1) \\ &= \frac{1}{1-F_1(0)}(f_2(y_2) - \frac{\partial}{\partial y_2}F(0, y_2)). \end{aligned}$$

Note that for notational convenience, in the above and in the subsequent paragraphs we have mainly used F in place of C to avoid cluttering the formulae. Result (2.4.1) cannot be used for discrete distributions; this trivially follows from the properties of $F_2(y_2)$ which is discrete and therefore discontinuous on the integers in its domain. Thus, $F_2(y_2)$ is not differentiable with respect to y_2. The same is true for $F(0, y_2)$ as it includes $F_2(y_2)$. Recall that $f_2(y_2)$ can be obtained as $f_2(y_2) = F_2(y_2) - F_2(y_2 - 1)$. Also, if $y_2 = 0$ then $f_2(0) = F_2(0)$ (because $F_2(-1)$ is not in the support of distributions considered and therefore will have probability equal to 0). In a bivariate context, if both margins are discrete then the joint copula pmf will be represented as (Nikoloulopoulos & Karlis, 2010; Karlis & Pedeli, 2013)

$$\begin{aligned} f(y_1, y_2) &= C(F_1(y_1), F_2(y_2); \theta) - C(F_1(y_1 - 1), F_2(y_2); \theta) \\ &\quad - C(F_1(y_1), F_2(y_2 - 1); \theta) + C(F_1(y_1 - 1), F_2(y_2 - 1); \theta). \qquad (2.4.2) \end{aligned}$$

Figure 2: Probability mass functions for the Poisson, negative binomial, Delaporte, Poisson inverse Gaussian and Sichel distributions. The parameter values have been chosen arbitrarily to show different shapes of the distributions. For Poisson, μ is 1, 2 or 5. For negative binomial and Poisson inverse Gaussian, μ and σ are (1, 1), (5, 2) and (30, 3). For Delaporte, μ, σ and ν are (1, 1, 0.1), (5, 2, 0.3) and (30, 3, 0.5). For Sichel, μ, σ and ν are (1, 1, 0.5), (5, 0.5, 1) and (8, 0.1, 3). Note that Delaporte and Sichel can have thinner or thicker tails depending on the choice of parameters. At the same time, the tails of Poisson are thinner than those of Delaporte and Sichel.



By taking finite differences of $F(y_1, y_2) = C(F_1(y_1), F_2(y_2); \theta)$ with respect to y_1, we obtain $C(F_1(y_1), F_2(y_2); \theta) - C(F_1(y_1 - 1), F_2(y_2); \theta)$. Again, calculating finite differences of each of these elements with respect to y_2 will yield (2.4.2). An analogical reasoning can be used in the current context where we have copulae with continuous-discrete margins.

In the second line of (2.4.1), the derivative with respect to y_2 is obtained by using finite differences. Thus,

$$\begin{aligned} f_{2|1}(y_2|y_1^* > 0) &= \frac{1}{1 - F_1(0)}\{F_2(y_2) - F_2(y_2 - 1)\} - \frac{1}{1 - F_1(0)}\{F(0, y_2) - F(0, y_2 - 1)\} \\ &= \frac{1}{1 - F_1(0)}\{f_2(y_2) - F(0, y_2) + F(0, y_2 - 1)\}. \end{aligned}$$

The likelihood therefore assumes the following form

$$L = \prod_0 F_1(0) \prod_1 \{f_2(y_2) - F(0, y_2) + F(0, y_2 - 1)\}$$

and the model log-likelihood is given by

$$\ell = \sum_0 \log F_1(0) + \sum_1 \log(f_2(y_2) - F(0, y_2) + F(0, y_2 - 1)). \tag{2.4.3}$$

The subscripts under capital sigmas indicate whether an observation is missing or not. The likelihood construction above does not require computationally intensive quadrature or simulations methods as, for instance, in Greene (1997) and Miranda & Rabe-Hesketh (2006). The exception is for the bivariate normal copula where Φ_2 is obtained by solving a double integral. Nevertheless, many efficient methods are available for calculating bivariate normal cdfs.

We would like to stress that first and second derivatives of ℓ with the respect to the model parameters can be obtained in a modular fashion. Let us denote the overall vector of the outcome equation parameters as $\boldsymbol{\delta}_2$. The general expression of the first derivative of (2.4.3) with respect to $\boldsymbol{\delta}_2$ is

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{\delta}_2} &= \sum_1 \frac{1}{f_2(y_2) - F(0,y_2) + F(0,y_2-1)} \times \\ &\left(\frac{\partial f_2(y_2)}{\partial \boldsymbol{\delta}_2} - \frac{\partial F(0,y_2)}{\partial F_2(y_2)}\frac{\partial F_2(y_2)}{\partial \boldsymbol{\delta}_2} + \frac{\partial F(0,y_2-1)}{\partial F_2(y_2-1)}\frac{\partial F_2(y_2-1)}{\partial \boldsymbol{\delta}_2}\right), \end{aligned} \tag{2.4.4}$$

This expression shows that there are two derivatives whose forms will depend on the chosen copula and three derivatives which are margin dependent. The main structure of (2.4.4) will, however, be unaffected by the specific choices made. This means that it will be relatively easy to extend our algorithm to other copula and discrete margin models. Full details on the score and Hessian equations are given in Supplementary Material 2.

3 Parameter estimation

It is common practice to estimate the parameters of models including regression spline components subject to some sort of penalization (see, e.g., Wood, 2006). This aims at avoiding overfitting which is likely to occur when employing flexible model specifications. To this end, a roughness penalty term is introduced in (2.4.3) (Ruppert et al., 2003; Wood, 2006). Define $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\delta}_1^\mathsf{T}, \boldsymbol{\delta}_2^\mathsf{T}, \sigma^*, \nu^*, \theta^*)$, where $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ are overall parameter vectors containing all regression coefficients associated with η_1 and η_2, σ^* and ν^* are outcome distribution specific parameters and θ^* is the association copula parameter. The star superscript is used when a parameter has been transformed so that its support is not bounded (see Appendix 1). Note that this parameter vector's definition is consistent with the Delaporte and Sichel distributions. For Poisson we would have $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\delta}_1^\mathsf{T}, \boldsymbol{\delta}_2^\mathsf{T}, \theta^*)$, whereas for Negative Binomial and Poisson inverse Gaussian $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\delta}_1^\mathsf{T}, \boldsymbol{\delta}_2^\mathsf{T}, \sigma^*, \theta^*)$. Each smooth $s_{vk_v}(u_{vk_v})$ has an associated penalty which can be expressed as $\boldsymbol{\alpha}_{vk_v}^\mathsf{T}\mathbf{S}_{vk_v}\boldsymbol{\alpha}_{vk_v}$, where $\mathbf{S}_{vk_v}$ is a positive semi-definite penalty matrix with known coefficients. This quadratic expression comes from writing explicitly $\int s''_{vk_v}(u_{vk_v})^2 du_{vk_v}$. Several definitions of $\mathbf{S}_{vk_v}$ are possible depending on type of spline basis employed (Wood, 2006). The penalized likelihood is

$$\ell_p(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) - \frac{1}{2}\boldsymbol{\delta}^\mathsf{T}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\delta}, \tag{3.0.5}$$

where $\mathbf{S}_{\boldsymbol{\lambda}} = \text{diag}(\mathbf{0}_{P_1}^\top, \lambda_{1k_1}\mathbf{S}_{1k_1}, ..., \lambda_{1K_1}\mathbf{S}_{1K_1}, \mathbf{0}_{P_2}^\top, \lambda_{2k_2}\mathbf{S}_{2k_2}, ..., \lambda_{2K_2}\mathbf{S}_{2K_2}, 0, 0, 0)$ for Delaporte and Sichel and λ_{vk_v} are smoothing parameters controlling the trade-off between fit and smoothness. If σ^*, ν^* and θ^* are modeled as functions of flexible linear predictors then $\boldsymbol{\delta}^\top = (\boldsymbol{\delta}_1^\top, \boldsymbol{\delta}_2^\top, \boldsymbol{\delta}_3^\top, \boldsymbol{\delta}_4^\top, \boldsymbol{\delta}_5^\top)$ and $\mathbf{S}_{\boldsymbol{\lambda}} = \text{diag}(\mathbf{0}_{P_1}^\top, \lambda_{1k_1}\mathbf{S}_{1k_1}, ..., \lambda_{1K_1}\mathbf{S}_{1K_1}, \mathbf{0}_{P_2}^\top, \lambda_{2k_2}\mathbf{S}_{2k_2}, ..., \lambda_{2K_2}\mathbf{S}_{2K_2}, \mathbf{0}_{P_3}^\top, \lambda_{3k_3}\mathbf{S}_{3k_3}, ..., \lambda_{3K_3}\mathbf{S}_{3K_3}, \mathbf{0}_{P_4}^\top,$ $\lambda_{4k_4}\mathbf{S}_{4k_4}, ..., \lambda_{4K_4}\mathbf{S}_{4K_4}, \mathbf{0}_{P_5}^\top, \lambda_{5k_5}\mathbf{S}_{5k_5}, ..., \lambda_{5K_5}\mathbf{S}_{5K_5})$, where the additional components have the obvious definitions. The estimation algorithm is structured in two main steps which are described in the next sections.

3.1 Estimating $\boldsymbol{\delta}$

We employ a trust region algorithm which proved effective in the context of flexible sample selection models (e.g., Marra & Radice, 2013a; Wojtyś, in press). The approach establishes a sphere around the current a^{th} iterate $\boldsymbol{\delta}^{[a]}$ within which the next iterate $\boldsymbol{\delta}^{[a+1]}$ is to be found. If for the candidate $\boldsymbol{\delta}^{[a+1]}$ the model function is not "close enough" to the objective, then the region is shrunk and a new candidate found. Because the candidate points never lie outside the region, the algorithm will not run too far from the current point. If the objective function is undefined or indeterminate at an iteration then the trust algorithm will reject the candidate $\boldsymbol{\delta}^{[a+1]}$, reduce the radius of the region and try to find $\boldsymbol{\delta}^{[a+1]}$ again (Nocedal & Wright, 2006). Thus, the trust region algorithms tend to be more stable and reliable than line search methods, especially for functions which are non-convex, ill-conditioned and have long plateaus (e.g., Braun, 2013). More details can be found in Supplementary Material 3.

Define $\boldsymbol{d}^{[a]} = \boldsymbol{\delta}^{[a+1]} - \boldsymbol{\delta}^{[a]}$. To obtain this at each iteration we seek the solution to the problem

$$\min_{d\in\mathbb{R}^P} m_a(\boldsymbol{d}^{[a]}) \quad = \quad -\{\ell_p(\boldsymbol{\delta}^{[a]}) + \boldsymbol{d}^{[a]\top}\mathbf{g}_p^{[a]} + \frac{1}{2}\boldsymbol{d}^{[a]\top}\mathbf{H}_p^{[a]}\boldsymbol{d}^{[a]}\} \qquad \text{subject to} \quad \|\boldsymbol{d}^{[a]}\| \leq \triangle^{[a]},$$

where $\mathbf{g}_p^{[a]} = \mathbf{g}^{[a]} - \mathbf{S}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}^{[a]}$ is the penalized score vector, $\mathbf{H}_p^{[a]} = \mathbf{H}^{[a]} - \mathbf{S}_{\hat{\boldsymbol{\lambda}}}$ is the penalized Hessian matrix, $\|\cdot\|$ represents the Euclidean norm, and $\triangle^{[a]}$ is the radius of the trust-region at iteration a. "Closeness" will be the benchmark for deciding about the trust region radius. The numerator of

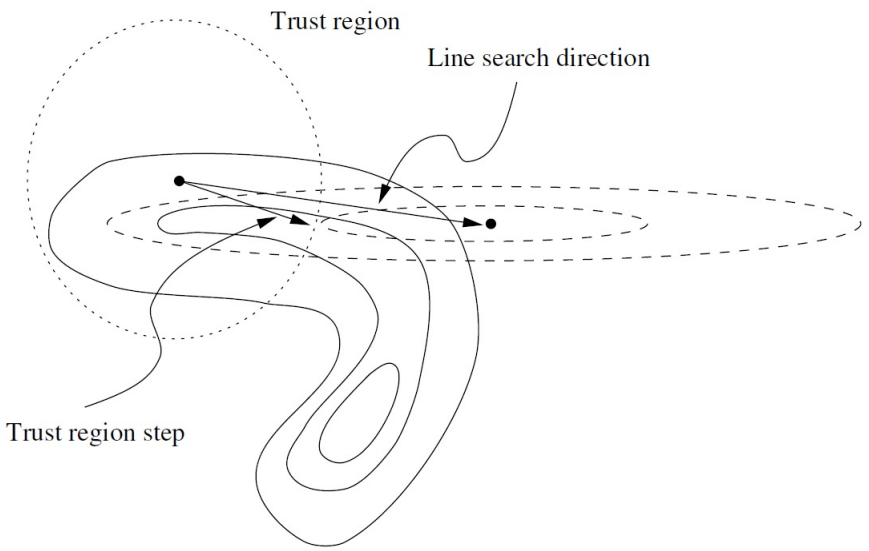
$$\rho^{[a]} \quad = \quad \frac{\ell_p(\boldsymbol{\delta}^{[a]}) - \ell_p(\boldsymbol{\delta}^{[a]} + \boldsymbol{d}^{[a]})}{m(\mathbf{0}) - m(\boldsymbol{d}^{[a]})},$$

represents the actual reduction in the objective function and the denominator represents the predicted one. The first terms of the numerator and denominator are evaluated at iteration a, i.e. $\boldsymbol{d}^{[a]} = \boldsymbol{\delta}^{[a]} - \boldsymbol{\delta}^{[a]} = \mathbf{0}$, whereas the second terms are evaluated at iteration $a+1$, i.e. $\boldsymbol{d}^{[a]} = \boldsymbol{\delta}^{[a+1]} - \boldsymbol{\delta}^{[a]}$. Note that close to the solution, the trust region typically behaves like a line search method. For illustrative purposes, Figure 3 compares the trust region approach to a line search method when using a two parameter likelihood function.

3.2 Smoothing parameter estimation

Automatic multiple smoothing parameter estimation is achieved by using an adaptation of the approach employed by Radice et al. (in press) in a similar context. Note that simultaneous optimization

Figure 3: Comparison of trust region algorithm against line search methods based on a two parameter likelihood function (figure from Nocedal & Wright, 2006). The current point lies in the upper left end part of the graph while the minimum point lies in the lower end of the valley. The quadratic model m_a is represented by the dashed elliptical contour lines. A line search method based on this model (longer arrow) would search along the step to the minimizer of m_a, allowing only for small reduction in the objective function. A trust-region method (shorter arrow) shifts to the minimizer of m_a within the dotted circle, which yields a more significant reduction in the function and a better step.



of $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ will lead to overfitting since the highest value of the penalized likelihood will be obtained when $\boldsymbol{\lambda} = \mathbf{0}$. In fact, $\boldsymbol{\lambda}$ should be chosen so that the estimated smooths are as close as possible to the true functions, as described below.

Following Yee & Wild (1996), problem (3.0.5) can be approximated as

$$\min_{\boldsymbol{\delta}\in\mathbb{R}^n} ||\sqrt{\mathbf{W}^{[a]}}(\mathbf{z}^{[a]} - \mathbf{X}\boldsymbol{\delta})||^2 + \boldsymbol{\delta}^\mathsf{T}\mathbf{S}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\delta}, \tag{3.2.1}$$

where design matrix $\mathbf{X}$ is of dimension $5n \times P$ and consists of $5 \times P$ sub-matrices $\mathbf{X}_i = \text{diag}(\mathbf{X}_{1i}^\mathsf{T}, \mathbf{X}_{2i}^\mathsf{T}, 1, 1, 1)$, $P = P_1 + K_1 + P_2 + K_2 + 3$ denotes the total number of parameters in the model, $\mathbf{X}_{1i} = (\boldsymbol{z}_i^\mathsf{T}, \mathbf{B}(u_{1k_1i})^\mathsf{T})^\mathsf{T}$ and $\mathbf{X}_{2i} = (\boldsymbol{x}_i^\mathsf{T}, \mathbf{B}(u_{2k_2i})^\mathsf{T})^\mathsf{T}$. Weight matrix is block diagonal, has dimension $5n \times 5n$ and is defined as $\mathbf{W} = -\partial^2\ell/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}$ where each block within $\mathbf{W}$ has dimension 5×5 and takes the form

$$\mathbf{W}_i = \frac{\partial^2\ell}{\partial\boldsymbol{\eta}_i\partial\boldsymbol{\eta}_i} = \begin{bmatrix} \frac{\partial^2\ell}{\partial\eta_{1i}\partial\eta_{1i}} & \cdot & \cdot & \cdot & \cdot \\ \frac{\partial^2\ell}{\partial\eta_{1i}\partial\eta_{2i}} & \frac{\partial^2\ell}{\partial\eta_{2i}\partial\eta_{2i}} & \cdot & \cdot & \cdot \\ \frac{\partial^2\ell}{\partial\eta_{1i}\partial\sigma^*} & \frac{\partial^2\ell}{\partial\eta_{2i}\partial\sigma^*} & \frac{\partial^2\ell}{\partial\sigma^{*2}} & \cdot & \cdot \\ \frac{\partial^2\ell}{\partial\eta_{1i}\partial\nu^*} & \frac{\partial^2\ell}{\partial\eta_{2i}\partial\nu^*} & \frac{\partial^2\ell}{\partial\sigma^*\partial\nu^*} & \frac{\partial^2\ell}{\partial\nu^{*2}} & \cdot \\ \frac{\partial^2\ell}{\partial\eta_{1i}\partial\theta^*} & \frac{\partial^2\ell}{\partial\eta_{2i}\partial\theta^*} & \frac{\partial^2\ell}{\partial\sigma^*\partial\theta^*} & \frac{\partial^2\ell}{\partial\nu^*\partial\theta^*} & \frac{\partial^2\ell}{\partial\theta^{*2}} \end{bmatrix},$$

with $\boldsymbol{\eta}_i$ defined as $(\eta_{1i}, \eta_{2i}, \sigma^*\nu^*, \theta^*)^\mathsf{T}$. Pseudo-data vector $\mathbf{z}^{[a]}$ is defined as $\mathbf{W}^{-1[a]}\mathbf{u}^{[a]} + \mathbf{X}\boldsymbol{\delta}^{[a]}$, where $\mathbf{u}^{[a]}$ is a vector consisting of n sub-vectors $\frac{\partial\ell}{\partial\boldsymbol{\eta_i}} = \left(\frac{\partial\ell}{\partial\eta_{1i}}, \frac{\partial\ell}{\partial\eta_{2i}}, \frac{\partial\ell}{\partial\sigma^*}, \frac{\partial\ell}{\partial\nu^*}, \frac{\partial\ell}{\partial\theta^*}\right)^\mathsf{T}$. Since σ^*, ν^* and θ^* are not specified as functions of predictors then $\sigma^* = \eta_{3i} = \eta_3$, $\nu^* = \eta_{4i} = \eta_4$ and $\theta^* = \eta_{5i} = \eta_5$ where

the η's contain only intercepts. (In some cases, we have suppressed the iteration index from some of the quantities described above to avoid clutter.) If σ^*, ν^* and θ^* are specified as functions of linear predictors then the sub-matrices $\mathbf{X}_i$ will be defined as $\mathbf{X}_i = \text{diag}(\mathbf{X}_{1i}^\mathsf{T}, \mathbf{X}_{2i}^\mathsf{T}, \mathbf{X}_{3i}^\mathsf{T}, \mathbf{X}_{4i}^\mathsf{T}, \mathbf{X}_{5i}^\mathsf{T})$, where $\mathbf{X}_{3i}$, $\mathbf{X}_{4i}$ and $\mathbf{X}_{5i}$ denote the covariate vectors corresponding to η_{3i}, η_{4i} and η_{5i}. The sub-vectors of $\mathbf{u}$ will be defined as $\frac{\partial \ell}{\partial \boldsymbol{\eta_i}} = \left(\frac{\partial \ell}{\partial \eta_{1i}}, \frac{\partial \ell}{\partial \eta_{2i}}, \frac{\partial \ell}{\partial \eta_{3i}}, \frac{\partial \ell}{\partial \eta_{4i}}, \frac{\partial \ell}{\partial \eta_{5i}}\right)^\mathsf{T}$, where $\frac{\partial \ell}{\partial \eta_{3i}}$, $\frac{\partial \ell}{\partial \eta_{4i}}$ and $\frac{\partial \ell}{\partial \eta_{5i}}$ are the partial derivatives of ℓ with respect to the linear predictors of σ_i^*, ν_i^* and θ_i^*. Finally, the derivatives in the sub-matrices of $\mathbf{W}$ which are related to σ_i^*, ν_i^* and θ_i^* will have to be re-written accordingly. The solution to (3.2.1) is

$$\tilde{\boldsymbol{\delta}} \;=\; \left(\mathbf{X}^\mathsf{T}\mathbf{W}^{[a]}\mathbf{X} + \mathbf{S}_{\hat{\boldsymbol{\lambda}}}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}^{[a]}\mathbf{z}^{[a]}.$$

We are now in a position to employ the approach discussed by Radice et al. (in press) for smoothing parameter estimation. This consists of minimizing

$$\mathcal{V}_u(\boldsymbol{\lambda}) \;=\; \frac{1}{n^*}||\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\tilde{\boldsymbol{\delta}})||^2 - 1 + \frac{2}{n^*}\,\text{tr}(\mathbf{A}), \tag{3.2.2}$$

where $n^* = 5n$, $\sqrt{\mathbf{W}}\mathbf{A} = \sqrt{\mathbf{W}}\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}$ is the hat matrix, and the weight matrix and pseudo-data vector are constructed from the current iterate for $\boldsymbol{\delta}$. The trace of $\mathbf{A}$ represents the effective degrees of freedom (*edf*) of the penalized model. Details on the derivation of the above criterion are given in Appendix 2.

The two steps (one described in the previous section and the other here) are iterated until convergence in an outer iteration fashion (O'Sullivan et al., 1986):

1. Using a starting parameter vector value $\boldsymbol{\delta}^{[a]}$ and fixing the smoothing parameter vector at $\boldsymbol{\lambda}^{[a]}$, find an estimate for $\boldsymbol{\delta}$:

$$\boldsymbol{\delta}^{[a+1]} \;=\; \text{arg}_{\boldsymbol{\delta}}\max \ell_p(\boldsymbol{\delta}).$$

2. Using $\boldsymbol{\delta}^{[a+1]}$ construct the working linear model components required in (3.2.2) and find an updated estimate for $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda}^{[a+1]} \;=\; \text{arg}_{\boldsymbol{\lambda}}\min \mathcal{V}_u(\boldsymbol{\lambda}).$$

The minimization in step 2 is performed using the computationally efficient and stable approach by Wood (2004). Note that because $\mathbf{W}$ is non-diagonal, the construction of the quantities required in (3.2.2) can be computationally costly as well as not feasible for large sample sizes. Here, dramatic computational savings are achieved by exploiting the sparse structure of $\mathbf{W}$ which makes it possible to set up the weighted pseudo-data vector and weighted design matrix using n small blocks of vectors and matrices of dimensions 5×1 and 5×5, respectively. Other numerical aspects are discussed in Appendix 1 whereas some arguments related to the asymptotic behavior of the proposed estimator

are given in Supplementary Material 4.

4 Confidence intervals and model selection

At convergence, the covariance matrix of $\hat{\boldsymbol{\delta}}$ can be easily shown to be equal to $\mathbf{H}_p^{-1}\mathbf{H}\mathbf{H}_p^{-1}$, where $\mathbf{H} = \mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X}$ and $\mathbf{H}_p = \mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \mathbf{S}_{\hat{\boldsymbol{\lambda}}}$. As shown by Marra & Wood (2012), however, this definition does not yield intervals with close-to-nominal coverage probabilities for smooth terms. They argue that the Bayesian version of the covariance matrix should be used instead because it contains both a bias and variance component in the frequentist sense. This approach is based on the result $N\left(\hat{\boldsymbol{\delta}}, \mathbf{H}_p^{-1}\right)$ and has been used successfully in a sample selection context (e.g., Marra & Radice, 2013a). Note that for a model containing only unpenalized terms, the Bayesian and frequentist covariance matrix will be the same. Confidence intervals for a generic smooth component can therefore be constructed using

$$N(\mathbf{B}(u_{ki})^\mathsf{T}\boldsymbol{\alpha}_k, \mathbf{B}(u_{ki})^\mathsf{T}\mathbf{V}_{\boldsymbol{\delta}_k}\mathbf{B}(u_{ki})), \tag{4.0.3}$$

where $\mathbf{V}_{\boldsymbol{\delta}_k}$ is the covariance matrix related to the k^{th} smooth. Using a Bayesian result for interval construction also means that intervals for non-linear functions of the model parameters (i.e., σ, ν, θ) can be easily obtained by posterior simulation as follows:

1. Draw n_{sim} random vectors from $N\left(\hat{\boldsymbol{\delta}}, \mathbf{H}_p^{-1}\right)$.
2. Calculate n_{sim} simulated realizations of the function of interest. For example, since $\sigma = \exp(\sigma^*)$, we have that $\boldsymbol{\sigma}^{sim} = (\sigma_1^{sim}, \sigma_2^{sim}, ..., \sigma_{n_{sim}}^{sim})$ where $\sigma_o^{sim} = \exp(\sigma_o^{*sim})$, $o = 1, 2, ..., n_{sim}$.
3. Using $\boldsymbol{\sigma}^{sim}$, calculate the lower, $\xi/2$, and upper, $1 - \xi/2$, quantiles. For 95% intervals, ξ is set to 0.05.

Result (4.0.3) is not appropriate for testing the null hypothesis that a smooth term is equal to zero (e.g., Ruppert et al., 2003). However, recently Wood (2013) developed a Wald-type test for smooth terms. This approach can be easily adapted to the context of bivariate equation system models as done, for instance, by Radice et. al (in press). Finally, variable (and or model) selection can be performed by employing commonly used techniques such as the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). In our context, these are defined as

$$\begin{aligned} \text{AIC} &= -2\ell(\hat{\boldsymbol{\delta}}) + 2edf \\ \text{BIC} &= -2\ell(\hat{\boldsymbol{\delta}}) + \log(n)edf, \end{aligned}$$

where $\ell(\hat{\boldsymbol{\delta}})$ is the likelihood of the sample selection model evaluated at the penalized parameter estimates and *edf* is defined in Section 3.2.

To assess visually the goodness of fit of a discrete marginal distribution, normal Q-Q plots of normalized quantile residuals can be employed. Such residuals are constructed by normalizing and

randomizing the residuals defined as $\hat{r} = \Phi^{-1}(u_i)$, where Φ^{-1} is the inverse cdf of a standard normal and u_i is a random value from the uniform distribution on $\left[F_2(y_{2i} - 1|\hat{\boldsymbol{\delta}}), F_2(y_{2i}|\hat{\boldsymbol{\delta}})\right]$ with y_{2i} denoting the observed response. Interpretation of Q-Q plots of such residuals will be the same as if our reference distribution was standard normal. Randomization consists of making the discrete distribution as if it was continuous (Smyth & Dunn, 1996). With regard to the selection equation, since the response is binary, residual analysis would not be informative unless residuals can be grouped somehow (e.g., Collett, 2002).

5 Simulations

To assess the empirical effectiveness of the proposed approach, we conducted a simulation study in which we fitted models to data generated under different scenarios. We considered estimating models with correct specification and misspecified copula and margin. We also fitted univariate models (ignoring non-random sample selection). The simulations were carried out using the `R` package `SemiParSampleSel` (Marra et al., 2016) which implements all developments discussed in the previous sections (see Appendix 3 for a brief description of the package).

The simulations were based on the following equations

$$\begin{aligned} Y_{1i}^* &= \gamma_0 + s_1(x_{1i}) + \gamma_1 x_{2i} + \gamma_2 x_{3i} + \epsilon_{1i}, \quad \epsilon_{1i} \sim N(0,1), \\ E(Y_{2i}) &= \exp(\beta_0 + s_2(x_{1i}) + \beta_1 x_{2i}), \end{aligned}$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^\mathsf{T} = (1.0, -2.0, 0.3)^\mathsf{T}$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^\mathsf{T} = (1.1, -1.9)^\mathsf{T}$,
$s_1(x_{1i}) = 0.4\left[-4 - (5.5x_{1i} - 2.9) + 3(4.5x_{1i} - 2.3)^2 - (4.5x_{1i} - 2.3)^3\right]$, $s_2(x_{1i}) = x_{1i}\sin(8x_{1i})$, and σ and Kendall's τ were fixed at 1 and 0.5. Following the approach of Marra & Wood (2012), we generated covariates from a multivariate normal distribution with zero mean vector, and covariance matrix with ones on the main diagonal and Pearson's correlation coefficient equal to 0.5. The covariates were then transformed using standard normal cdfs. s_1 and s_2 were modeled using thin plate regression splines with 10 bases and penalties based on second order derivatives. Unconditional expected mean values were calculated and the parameters μ and σ inserted into `mvdc()` from the `copula` package which delivered random outcome responses from a given copula. Using the above specification, at each replicate around 42% of the observations were selected for the outcome equation. The `R` code used to generate the data can be found in Supplementary Material 5.

The simulation study aimed at investigating the performance of the proposed methodology under correct specification, copula misspecification and margin misspecification. We also assessed the coverage of the smooth functions in the model. To keep the study feasible, we mainly considered the following situations:

1. *Copula misspecification.* The data generating process (DGP) was Clayton copula with negative binomial margin. We ran two batches of simulations with 250 replicates, and 3000 and 6000 observations each. For each replicate, we fitted models under correct specification and under

misspecification where the normal, Frank, Joe and Gumbel copulae were employed. We also fitted univariate models ignoring non-random sample selection.

2. *Margin misspecification.* The same setup as above was used for the number of replicates and observations. The DGP was negative binomial with Frank copula and the estimated models included the correctly specified one as well as models using Poisson and Poisson inverse Gaussian margins.

3. *Coverage of s_1 and s_2.* The number of replicates was increased to 1000 and for each replicate we generated data and fitted the correctly specified model. The sample size was 3000, the copulae used were Clayton, Frank, normal, Joe and Gumbel, and the margin was negative binomial.

The specification used to generate the data contains an exclusion restriction, x_3, which typically helps with empirical identification. This variable has an impact on the selection process but not on the response of interest. If both equations include the same covariates then the model is still theoretically identified, although estimation results may depend more heavily on functional form assumptions (see Puhani (2000) for a discussion of this aspect in a related context). In this regard, we conducted a round of simulations without exclusion restriction. The results were very close to those obtained in the presence of an exclusion restriction and are shown in Tables 3 and 4 (more results are available upon request). We also tried different configurations of parameter values but the main conclusions did not change. It is worth pointing out that it is difficult (if not impractical) to simulate the highly complex processes that likely underlie the relation between a selection process and a response of interest. Therefore, although the simulation results suggest that an exclusion restriction may not be required for empirical identification, when an exclusion restriction is not imposed we cannot rule out the possibility that under scenarios that are different from those considered here the model may not correct satisfactorily for selection bias.

Tables 3 and 4 show simulation results related to points 1. and 2. above. The second and third columns show the percentages of times a model was preferred over its competitors by AIC and BIC. In all scenarios, the true model was preferred most of the times. We also calculated the root mean squared error (RMSE) of $\hat{s}_2$, $\hat{\beta}_1$, $\hat{\sigma}$ and Kendall's $\hat{\tau}$. For the last three parameters, we also computed the percentage bias. The true model has the lowest RMSE and percentage bias. In the case of copula misspecification (Table 3), the closest counterpart is the negative binomial model with Frank copula which, overall, produced results that are very close to the reference model and that are substantially better than those obtained using univariate models. Selecting a copula which is not supported by AIC and BIC (in this case, Joe and Gumbel) produced less reliable estimates especially for β_1 where the univariate models gave less biased estimates for this parameter. This suggests that the choice of copula matters and that criteria such as AIC and BIC can help making the most appropriate decision. Regarding margin misspecification (Table 4), similar conclusions can be drawn; misspecifying the margin can have important consequences on parameter estimation and AIC and BIC can help choosing the most appropriate margin.

Figure 4 depicts the smooth function estimates for all replicates for the scenarios with negative

binomial margin and some of the copulae considered in the simulations. The upper and lower plots display the results for the outcome equation when using the sample selection and univariate models. The proposed modeling approach recovers the underlying true curves well. The results from the univariate models show the detrimental impact that ignoring non-random sample selection has on function estimates. We also calculated point-wise coverage probabilities for s_2 as done in Marra & Wood (2012). The intervals produced close-to-nominal coverage probabilities (e.g., coverages between 93% and 96% for a nominal level of 95%) with the best results associated with the correctly specified model.

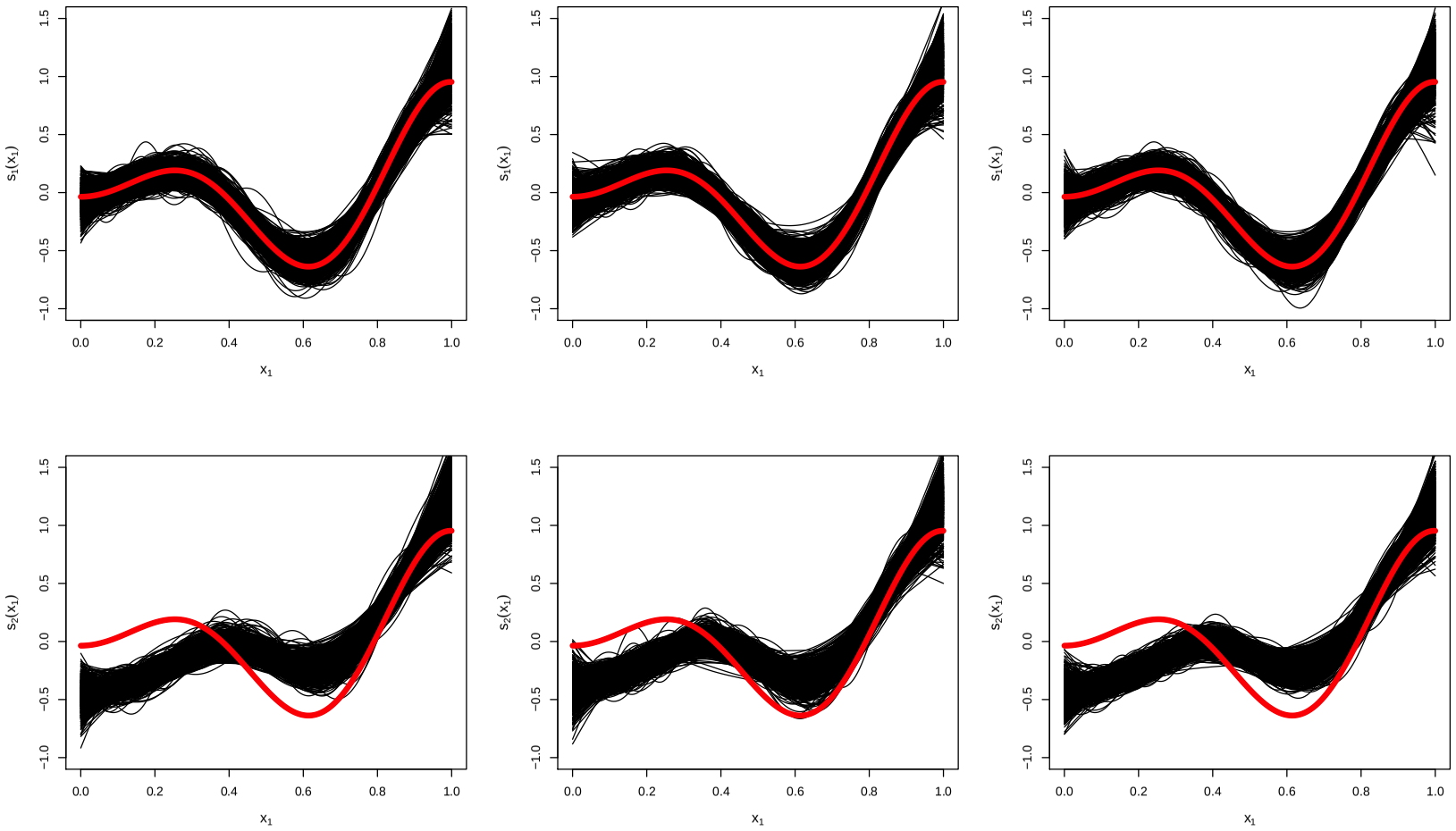
No ER $n = 3000$									
	AIC	**BIC**	**RMSE** $\hat{s}_2(x_1)$	**RMSE** $\hat{\beta}_1$	**% bias** $\hat{\beta}_1$	**RMSE** $\hat{\sigma}$	**% bias** $\hat{\sigma}$	**RMSE** $\hat{\tau}$	**% bias** $\hat{\tau}$
Univariate	0.00	0.00	0.251	0.204	-9.5	0.295	-29.1	-	-
DGP	93.17	95.58	0.085	0.103	0.5	0.069	-1.3	0.043	-0.9
Normal	1.61	1.20	0.091	0.225	10.2	0.108	-8.5	0.075	-12.1
Frank	5.22	3.21	0.091	0.148	5.4	0.073	-2.2	0.047	-4.2
Joe	0.00	0.00	0.149	0.300	9.5	0.205	-17.8	0.289	-48.2
Gumbel	0.00	0.00	0.107	0.288	13.2	0.138	-11.2	0.133	-20.4
No ER $n = 6000$									
	AIC	**BIC**	**RMSE** $\hat{s}_2(x_1)$	**RMSE** $\hat{\beta}_1$	**% bias** $\hat{\beta}_1$	**RMSE** $\hat{\sigma}$	**% bias** $\hat{\sigma}$	**RMSE** $\hat{\tau}$	**% bias** $\hat{\tau}$
Univariate	0.00	0.00	0.245	0.197	-9.5	0.292	-29.0	-	-
DGP	96.00	97.20	0.070	0.085	0.5	0.055	-0.8	0.037	-0.2
Normal	0.40	0.00	0.080	0.220	10.4	0.097	-8.1	0.070	-11.6
Frank	3.60	2.80	0.082	0.138	5.6	0.060	-1.7	0.040	-3.6
Joe	0.00	0.00	0.134	0.294	9.8	0.198	-17.8	0.283	-48.0
Gumbel	0.00	0.00	0.094	0.286	13.8	0.123	-10.6	0.112	-18.9
ER $n = 3000$									
	AIC	**BIC**	**RMSE** $\hat{s}_2(x_1)$	**RMSE** $\hat{\beta}_1$	**% bias** $\hat{\beta}_1$	**RMSE** $\hat{\sigma}$	**% bias** $\hat{\sigma}$	**RMSE** $\hat{\tau}$	**% bias** $\hat{\tau}$
Univariate	0.00	0.00	0.241	0.231	-11.1	0.298	-29.5	-	-
DGP	97.2	99.2	0.085	0.101	-0.3	0.071	-1.6	0.047	-1.6
Normal	0.00	0.00	0.088	0.225	10.1	0.117	-9.6	0.084	-13.7
Frank	2.8	0.8	0.087	0.149	5.5	0.077	-3.2	0.052	-5.0
Joe	0.00	0.00	0.151	0.307	7.1	0.228	-20.7	0.33	-56.5
Gumbel	0.00	0.00	0.099	0.296	13.4	0.150	-12.8	0.147	-23.2
ER $n = 6000$									
	AIC	**BIC**	**RMSE** $\hat{s}_2(x_1)$	**RMSE** $\hat{\beta}_1$	**% bias** $\hat{\beta}_1$	**RMSE** $\hat{\sigma}$	**% bias** $\hat{\sigma}$	**RMSE** $\hat{\tau}$	**% bias** $\hat{\tau}$
Univariate	0.00	0.00	0.234	0.219	-10.9	0.296	-29.5	-	-
DGP	98.40	99.20	0.065	0.077	-0.1	0.054	-1.2	0.036	-1.2
Normal	0.00	0.00	0.069	0.219	10.4	0.107	-9.4	0.079	-13.8
Frank	1.60	0.80	0.070	0.139	5.8	0.061	-2.8	0.043	-4.8
Joe	0.00	0.00	0.138	0.287	5.9	0.234	-22.0	0.342	-60.9
Gumbel	0.00	0.00	0.077	0.290	14.1	0.139	-12.7	0.127	-22.6

Table 3: Copula misspecification scenario. The DGP was Clayton copula with negative binomial margin. The number of replicates was 250. For each replicate, we estimated six models with the same margin but different dependence structure: univariate, DGP (i.e. Clayton), normal, Frank, Joe and Gumbel. The first and second columns indicates the percentages of times AIC and BIC were the lowest. The remaining columns show root mean squared error and percentage bias of the quantities of interest. Four rounds of simulations were conducted - with exclusion restriction (ER) and without (no ER), with total number of observations equal to 3000 and 6000.

No ER $n = 3000$							
	AIC	**BIC**	**RMSE** $\hat{s}_2(x_1)$	**RMSE** $\hat{\beta}_1$	**% bias** $\hat{\beta}_1$	**RMSE** $\hat{\tau}$	**% bias** $\hat{\tau}$
Univariate	0.00	0.00	0.299	0.287	-14.1	-	-
DGP	99.6	99.6	0.084	0.112	-0.2	0.036	0.0
Poisson	0.00	0.00	0.323	0.396	-20.1	0.435	-86.9
Poisson Inverse	0.04	0.04	0.010	0.119	-2.0	0.056	-8.5
No ER $n = 6000$							
	AIC	**BIC**	**RMSE** $\hat{s}_2(x_1)$	**RMSE** $\hat{\beta}_1$	**% bias** $\hat{\beta}_1$	**RMSE** $\hat{\tau}$	**% bias** $\hat{\tau}$
Univariate	0.00	0.00	0.291	0.274	-13.99	-	-
DGP	100.00	100.00	0.066	0.067	-0.15	0.028	-0.1
Poisson	0.00	0.00	0.316	0.385	-19.95	0.433	-86.6
Poisson Inverse Gau	0.00	0.00	0.082	0.077	-1.86	0.051	-8.6
ER $n = 3000$							
	AIC	**BIC**	**RMSE** $\hat{s}_2(x_1)$	**RMSE** $\hat{\beta}_1$	**% bias** $\hat{\beta}_1$	**RMSE** $\hat{\tau}$	**% bias** $\hat{\tau}$
Univariate	0.00	0.00	0.280	0.323	-16.3	-	-
DGP	100.00	100.00	0.083	0.097	-0.7	0.042	-0.1
Poisson	0.00	0.00	0.296	0.433	-22.3	0.451	-90.1
Poisson Inverse Gaus	0.00	0.00	0.097	0.109	-2.6	0.063	-9.5
ER $n = 6000$							
	AIC	**BIC**	**RMSE** $\hat{s}_2(x_1)$	**RMSE** $\hat{\beta}_1$	**% bias** $\hat{\beta}_1$	**RMSE** $\hat{\tau}$	**% bias** $\hat{\tau}$
Univariate	0.00	0.00	0.274	0.311	-16.1	-	-
DGP	100.00	100.00	0.063	0.066	-0.3	0.029	-0.5
Poisson	0.00	0.00	0.292	0.425	-22.2	0.449	-89.8
Poisson Inverse Gaus	0.00	0.00	0.080	0.077	-2.15	0.057	-9.9

Table 4: Margin misspecification scenario. The DGP was Frank copula with negative binomial margin. The number of replicates was 250. For each replicate, we estimated four models with the same copula but different margins: univariate model with negative binomial margin, DGP (i.e. negative binomial margin), Poisson and Poisson Inverse Gaussian. For further details see the caption of Table 3.

Figure 4: Simulation results for the negative binomial margin scenarios with normal, Clayton and Frank copulae, respectively. The plots display the results for the outcome equation when using the sample selection (first row) and univariate models (second row). The red thick curves depict the true functions. The black lines represent the smooth estimates obtained for all 1000 simulation replicates.

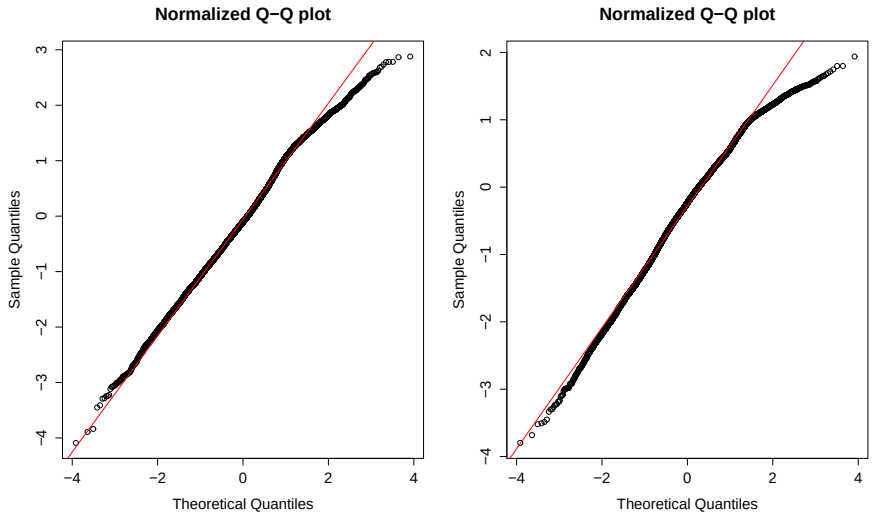


6 Empirical illustration

The data set on outpatient visits originates from the 2001 USVA Survey. As suggested by Lahiri & Xing (2004), it is of interest to evaluate the impact of patient's features on the number of visits in outpatient facilities and the predicted average number of visits. Outpatients are individuals who do not stay overnight in the hospital and attend, for instance, doctor visits and medical tests. Here, non-random sample selection arises since some potential patients may not be treated based on individual level characteristics, some of which are not observable. If a patient is treated then there is the option of choosing between VA and non-VA facilities and the number of times a treatment was received is recorded (Lahiri & Xing, 2004). In this study, possible unobservables are unidentified health shocks that might increase the number of hospital visits (e.g., car accidents), attitude towards health risks (e.g., smoking and drinking) and life style (e.g., sport activity); for analogical examples see Manning et al. (1988) and Trivedi & Zimmer (2007, p. 76).

The dataset contains 14,140 and the regressors used in the analysis are listed in Table 5. The variables of interest are number of VA outpatient visits (`NUMVAOUT`) and number of other outpatient visits (`NUMOTH`). For both responses, the preferred discrete marginal distribution is Poisson inverse Gaussian. This was determined by modeling the data using all discrete distributions available and then checking the QQ-plots of normalized and randomized residuals built as described in Section 4. In practice, this was achieved post estimation using `post.check()` in `SemiParSampleSel` which uses sample selection model estimates. Figure 5 shows the results; despite deviations from the reference lines, Poisson inverse Gaussian was the best-fitting distribution out of all distributions listed in Table 2.

Figure 5: QQ-plots (based on normalized and randomized residuals) for the outcome margin of the sample selection models based on the Poisson inverse Gaussian distribution. The two outcomes are NUMVAOUT (left) and NUMOTH (right).



For each response, the Frank and Gaussian copulae were fitted first to check whether the association between the selection and outcome equations was positive or negative. In this case it was

Variable	Description
VETSAGE	Age of veteran in years
WHITE	White race
MEDICAID	Covered by Medicaid/Medi-Cal
MEDICARE	Covered by Medicare
EXTSEX	Individual's gender
	Last year treatment received for...
HIGHBP	high blood pressure
LUNG	lung trouble
HEAR	a hearing condition requiring a hearing aid
ENT	other ear, throat, nose condition
EYE	eye/vision problem including needing glasses
CANCER	cancer
HEART	heart trouble
STROKE	a stroke
KIDNEY	kidney or bladder trouble
RHEUM	arthritis or rheumatism
LIVER	hepatitis C or other liver disease
HIV	an immune deficiency disease like HIV/AIDS
DIABETES	diabetes requiring insulin or diet treatment
STOMACH	stomach or digestive disorder
CHRONIC	severe chronic pain
DRUGS	drug abuse or alcoholism
PTSD	post-traumatic stress disorder
MENTAL	other mental or emotional problems
INJURY	an accident-related injury
TXOTH	any other serious condition

Table 5: Veterans Administration description of covariates. VETSAGE is a continuous variable. Regressors indicating a medical condition are binary.

negative, therefore the additional copulae considered were Clayton, Joe and Gumbel rotated by 90 and 270 degrees. The linear predictors of the selection and outcome equations, for both responses of interest, were specified following the findings of Lahiri & Xing (2004):

$$\begin{aligned}\eta_1 \;=\;& \gamma_0 + s(\texttt{VETSAGE}) + \gamma_1\texttt{EXTSEX} + \\ & \gamma_2\texttt{HIGHBP} + \gamma_3\texttt{LUNG} + \gamma_4\texttt{HEAR} + \\ & \gamma_5\texttt{ENT} + \gamma_6\texttt{EYE} + \gamma_7\texttt{CANCER} + \gamma_8\texttt{HEART} + \\ & \gamma_9\texttt{STROKE} + \gamma_{10}\texttt{KIDNEY} + \gamma_{11}\texttt{RHEUM} + \gamma_{12}\texttt{LIVER} + \\ & \gamma_{13}\texttt{HIV} + \gamma_{14}\texttt{DIABETES} + \gamma_{15}\texttt{STOMACH} + \gamma_{16}\texttt{CHRONIC} + \\ & \gamma_{17}\texttt{DRUGS} + \gamma_{18}\texttt{PTSD} + \gamma_{19}\texttt{MENTAL} + \gamma_{20}\texttt{INJURY} + \\ & \gamma_{21}\texttt{TXOTH},\end{aligned}$$

and

$$\begin{aligned}\eta_2 \;=\;& \beta_0 + s(\texttt{VETSAGE}) + \beta_1\texttt{WHITE} + \\ & \beta_2\texttt{MEDICAID} + \beta_3\texttt{MEDICARE} + \beta_4\texttt{HIGHBP} + \beta_5\texttt{LUNG} + \\ & \beta_6\texttt{HEAR} + \beta_7\texttt{ENT} + \beta_8\texttt{EYE} + \beta_9\texttt{CANCER} + \\ & \beta_{10}\texttt{HEART} + \beta_{11}\texttt{STROKE} + \beta_{12}\texttt{KIDNEY} + \beta_{13}\texttt{RHEUM} + \\ & \beta_{14}\texttt{LIVER} + \beta_{15}\texttt{HIV} + \beta_{16}\texttt{DIABETES} + \beta_{17}\texttt{STOMACH} + \\ & \beta_{18}\texttt{CHRONIC} + \beta_{19}\texttt{DRUGS} + \beta_{20}\texttt{PTSD} + \beta_{21}\texttt{MENTAL} + \\ & \beta_{22}\texttt{INJURY} + \beta_{23}\texttt{TXOTH}.\end{aligned}$$

where the effects of `VETSAGE` were modeled using thin plate regression splines with 10 bases and penalties based on second order derivatives. Table 6 shows the AIC and BIC for all models. The lowest AIC and BIC values for the selection model in which `NUMVAOUT` was used as outcome correspond to Joe 270. For the model that employed `NUMOTH` instead, the copula with lowest AIC and BIC is Gumbel 90 followed by Normal.

	NUMVAOUT		NUMOTH	
Model	AIC	BIC	AIC	BIC
Normal	44586	45034	66518	66947
Frank	44571	45019	66681	67110
Clayton 90	44515	44962	66666	67096
Clayton 270	44608	45055	66538	66966
Gumbel 90	44606	45045	66516	66944
Gumbel 270	44569	45018	66553	66981
Joe 90	44606	45045	66545	66973
Joe 270	44508	44955	66665	67095

Table 6: Akaike and Bayesian information criteria (AIC and BIC) for the sample selection models based on Normal, Frank, Clayton, Joe and Gumbel, and rotations by 90 and 270 degrees.

Table 7 shows the average predictions and association parameters for the Normal, Frank and preferred rotated copulae (details on prediction calculations are given in Supplementary Material 7). The values in brackets indicate 95% intervals. In all cases, the estimates and intervals for θ and τ suggest that non-random sample selection is present. The predicted values for the `NUMVAOUT` response do not vary substantially when looking at the results from the univariate model (ignoring sample selection) and those from the copula models. This is not surprising given that the estimated associations indicate that selection on unobservables is not strong. On the other hand, the average prediction for `NUMOTH` from the univariate model is significantly lower than those obtained from the selection models. The average predictions also differ among the selection models. For instance, Clayton 270 has a higher average prediction than Frank does. This shows that the choice of copula can have an impact on the results of interest and that the proposed approach allows one to assess the

sensitivity of results to different modeling assumptions. It is worth noting that the rotated copula models produce the same predictions. This is because they have very similar shapes.

NUMVAOUT			
Model	$\bar{y}$	$\hat{\theta}$	$\hat{\tau}$
Univariate	1.95	-	-
	(1.78, 2.12)	-	-
Normal	2.19	-0.27	-0.18
	(1.95, 2.42)	(-0.38, -0.17)	(-0.25, -0.10)
Frank	2.06	-1.73	-0.19
	(1.88, 2.23)	(-2.26, -1.20)	(-0.24, -0.13)
Clayton 90	2.05	-1.19	-0.37
	(1.89, 2.22)	(-1.49, -0.94)	(-0.43, -0.32)
Joe 270	2.05	-2.19	-0.39
	(1.89, 2.21)	(-2.46, -1.96)	(-0.44, -0.35)

NUMOTH			
Model	$\bar{y}$	$\hat{\theta}$	τ
Univariate	3.64	-	-
	(3.57, 3.70)	-	-
Normal	6.13	-0.88	-0.69
	(5.85, 6.42)	(-0.90, -0.86)	(-0.71, -0.66)
Frank	5.16	-7.30	-0.57
	(4.83, 5.49)	(-8.73, -5.92)	(-0.63, -0.51)
Clayton 270	6.18	-2.29	-0.53
	(5.86, 6.50)	(-2.64, -1.99)	(-0.57, -0.50)
Gumbel 90	6.18	-2.66	-0.62
	(5.88, 6.48)	(-2.92, -2.45)	(-0.66, -0.59)

Table 7: Estimates and 95% intervals (in brackets) for the average predictions and association parameters from the univariate model and the selection models based on the normal, Frank, Clayton, Joe and Gumbel copulae with Poisson inverse Gaussian discrete outcome margin.

Figure 6 shows the smooth term plots for the selection and outcome equations of the preferred model for `NUMOTH` (Poisson inverse Gaussian model with Gumbel 90). Further plots are reported in Supplementary Material 8; note that in this case, the shape of the curve estimates does not seem to be sensitive to the choice of copula. The selection equation smooth estimate suggests that the probability of having an outpatient VA visit increases as age increases up to about 65 years of age, after which it starts decreasing gradually. The shape of the outcome equation smooth indicates that the number of visits decreases with age and then increases. That is, patients only require non-VA services once they have reached a certain age; until then they rely on VA facilities. The smooth outcome equation estimate for the univariate model is displayed in Figure 7. The shapes are overall similar, however the bottom of the valley is shifted from 62 to 59 years. Also, note that the smooth from the univariate model first increases and then decreases with age, after which it increases again up until the age of 80.

Figure 6: Selection (left) and outcome (right) smooth function estimates from the Poisson inverse Gaussian model with Gumbel90 copula for NUMOTH.

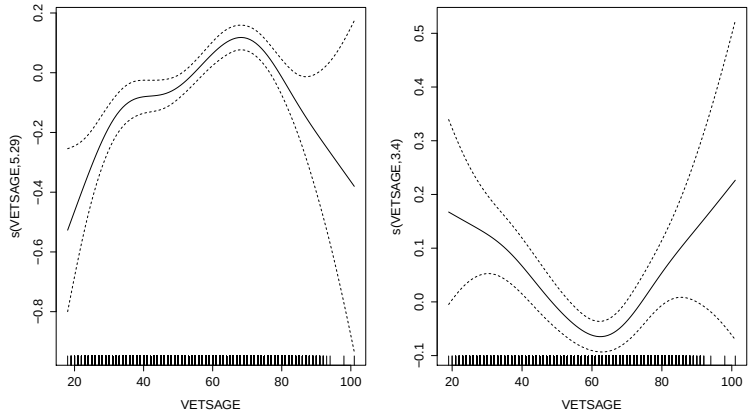


Figure 7: Smooth function estimates from the univariate Poisson inverse Gaussian model for NUMOTH.

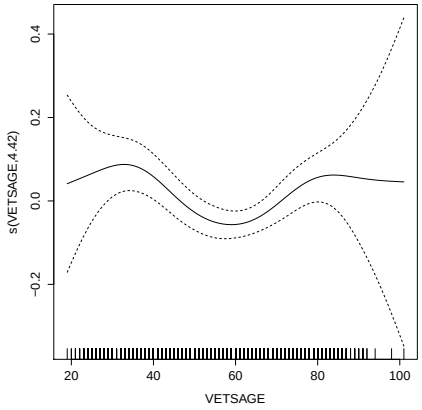


Table 8 shows the parametric effects for the outcome equation of the univariate and preferred selection models for NUMOTH. For instance, on average white patients more frequently use non-VA facilities than non-whites. This is consistent with Lahiri & Xing (2004). Similarly, Medicaid insurance holders are more likely to use non-VA facilities. On the other hand, there is no significant difference in terms of non-VA facility usage between Medicare insurance holders and non-holders. Hence, it seems that Medicare insurance holders are not discriminated against other insurers. Interestingly, patients with injuries are less likely to access non-VA facilities. Also, patients who have high blood pressure tend use non-VA outpatient facilities more frequently. The largest differences between both models can be observed for the intercepts and for the scale parameter σ. Thus, the discrepancy between the selection and univariate model predictions in Table 7 can be mostly attributed to the intercepts. This has also been found in other studies (McGovern et. al, 2015).

	Selection model (outcome equation)	Univariate model		Selection model (outcome equation)	Univariate model
Intercept	2.25	3.36	KIDNEY	-0.04	-0.06
	(0.46)	(0.41)		(0.03)	(0.03)
WHITE	0.14	0.17	RHEUM	0.04	-0.04
	(0.03)	(0.03)		(0.03)	(0.02)
MEDICAID	0.09	0.08	LIVER	0.01	-0.04
	(0.05)	(0.05)		(0.07)	(0.06)
MEDICARE	-0.02	-0.03	HIV	-0.41	-0.39
	(0.03)	(0.04)		(0.19)	(0.16)
HIGHBP	0.10	-0.04	DIABETES	-0.10	-0.13
	(0.02)	(0.02)		(0.03)	(0.03)
LUNG	-0.02	-0.06	STOMACH	-0.01	-0.09
	(0.03)	(0.03)		(0.03)	(0.03)
HEAR	0.04	0.07	CHRONIC	0.03	-0.05
	(0.03)	(0.03)		(0.03)	(0.03)
ENT	-0.04	-0.17	DRUGS	0.38	0.44
	(0.03)	(0.03)		(0.10)	(0.09)
EYE	0.03	-0.08	PTSD	0.11	0.15
	(0.02)	(0.02)		(0.05)	(0.05)
CANCER	-0.11	-0.18	MENTAL	0.05	-0.004
	(0.04)	(0.03)		(0.04)	(0.04)
HEART	-0.12	-0.18	INJURY	-0.17	-0.25
	(0.04)	(0.02)		(0.04)	(0.03)
STROKE	-0.08	0.002	TXOTH	-0.02	-0.17
	(0.06)	(0.05)		(0.03)	(0.03)
			σ	1.53	0.74
				(0.07)	(0.01)

Table 8: Estimates of parametric effects for the outcome equation of the univariate and preferred sample selection models for `NUMOTH`. The values in brackets indicate standard errors.

In summary, the above analysis shows that accounting for selection on unobservables can have a substantial impact on empirical results and that the proposed approach is a useful device to account for systematic missingness and assess the sensitivity of results to different modeling assumptions.

7 Conclusions

We have introduced a flexible copula-based regression framework to model count data suffering from non-random sample selection. The proposed approach allows for the use of several dependence structures, potentially any discrete outcome margin and for flexible covariate effects. The method also allows one to model distribution specific parameters as functions of flexible linear predictors. All developments are implemented in the `SemiParSampleSel R` package. The modeling framework has been illustrated in simulation and using 2001 USVA data.

The number of discrete margins presented in this paper is not exhaustive and a next release of `SemiParSampleSel` will incorporate more distributions such as beta binomial and zero inflated Poisson. It is worth mentioning that `SemiParSampleSel` allows for several types of smooth components. For instance, random effects smooths, Gaussian Markov random field smoothers, varying coefficient models (obtained, e.g., by multiplying one or more smooth components by some predictor), and smooth functions of two or more continuous covariates can also be employed for modeling.

An interesting extension would be to consider trivariate system models, controlling for the endogeneity of a treatment variable and for non-random sample selection in the outcome. In the context of the application of this paper, such a framework could be used for Medicare and Medicaid, which can arguably suffer from endogeneity. The proposed approach could also be extended to include two parameter copulae (see, e.g., Brechmann & Schepsmeier, 2013). This would lead to a better control of tail-dependence, although the association parameters may lose their interpretation. Since marginal distributions for the selection equation other than Gaussian may be plausible in applications, another venue for future research could be to employ skew probit links as derived from the standard skew-normal distribution by Azzalini (1985). As pointed out by Azzalini and Arellano-Valle (2013), in a considerably simpler context, the introduction of a parameter which regulates the distribution's skewness has very attractive properties from the probability point of view. However, a practical problem in applications is the possibility that the maximum likelihood estimate of the skewness parameter diverges, an issue which needs to be carefully addressed.

Acknowledgements

The second author would like to thank University College London for supporting this work with the University College London Impact Stipendship 2012-2015. The authors would also like to thank Christian Hennig, Rosalba Radice and Dr Małgorzata Wojtyś for many useful discussions on the preliminary versions of this work.

Supplementary Material

The on-line Supplementary Material contains seven sections. Section 1 discusses penalized regression splines and provides some examples. Section 2 provides information about the structure of the score vector, Hessian as well as some specific copula derivatives. Section 3 explains the trust region algorithm in more detail. Some arguments on the asymptotic properties of the proposed estimator can be found in Section 4. Section 5 reports the code used to simulate the data for the simulation study. In Section 6, model predictions are derived. Finally, Section 7 provides further results drawn from the empirical illustration.

The data set `public.sas7bdat` used for illustrating the framework is available upon request from the United States Veterans Administration web portal:

`http://www.va.gov/vetdata/`

References

[1] Azzalini, A., 1985. A class of distributions which includes the normal one. Scandinavian Journal of Statistics. 12, 171-178

[2] Azzalini, A., Arellano-Valle, R.B., 2013. Maximum penalized likelihood estimation for skew-normal and skew-t distributions. Journal of Statistical Planning and Inference 143, 419-433.

[3] Braun, M., 2014. trustOptim: An R Package for Trust Region Optimization with Sparse Hessians. Journal of Statistical Software, 60(4), 1–16.

[4] Brechmann, E.C., Schepsmeier, U., 2013. Modeling dependence with C- and D-vine copulas: The R-package CDVine. Journal of Statistical Software 52(3), 1–27.

[5] Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics: methods and applications. Cambridge university press.

[6] Cameron, A.C., Trivedi, P.K., 2013. Regression analysis of count data. Cambridge university press.

[7] Cameron, A.C., Li, T., Trivedi, P.K., Zimmer, D.M., 2004. Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. The Econometrics Journal 7(2), 566–584.

[8] Chib, S., Greenberg, E., Jeliazkov, I., 2009. Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection. Journal of Computational and Graphical Statistics, 18(2), 321–348.

[9] Collett, D., 2002. Modelling Binary Data. London: Chapman & Hall/CRC Texts in Statistical Science.

[10] Genest, C., Neslehova, J., 2007. A primer on copulas for count data. Astin Bulletin 37(2), 475–515.

[11] Genius, M., Strazzera, E., 2008. Applying the copula approach to sample selection modelling. Applied Economics 40(11), 1443–1455.

[12] Greene, W.H., 1997. FIML Estimation of Sample Selection Models for Count Data. Leonard N. Stern School of Business, New York.

[13] Gronau, R., 1974. Wage comparisons: A selectivity bias. Journal of Political Economy 82, 1119–1143.

[14] Hasebe, T., Vijverberg, W., 2012. A flexible sample selection model: A GTL-copula approach. IZA Discussion Paper, IZA, Bonn.

[15] Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–162.

[16] Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5(4), 475–492.

[17] Humphreys, B.R., 2013. Dealing With Zeros in Economic Data, Working paper, University of Alberta, Department of Economics.

[18] Karlis, D., Pedeli, X., 2013. Flexible Bivariate INAR (1) Processes Using Copulas. Communications in Statistics - Theory and Methods 42(4), 723–740.

[19] Lahiri, K., Xing, G., 2004. An econometric analysis of veterans' health care utilization using two-part models. Empirical Economics 29, 431–449.

[20] Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34(1), 1–14.

[21] Lewis, H.G., 1974. Comments on selectivity biases in wage comparisons. Journal of Political Economy 82(6), 1145–1155.

[22] Li, P., 2011. Estimation of sample selection models with two selection mechanisms. Computational Statistics and Data Analysis 55(2), 1099–1108.

[23] Liu, M., Kasteridis, P., Yen, S.T., 2012. Who are consuming food away from home and where? Results from the Consumer Expenditure Surveys. European Review of Agricultural Economics 5(1), 191–213.

[24] Manning, W.G., Newhouse, J.P, Duan, N., Keeler, E., Benjamin, B., Leibowitz, A., Zwanziger, J., 1988. Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment. RAND.

[25] Marchenko, Y.V., Genton, M.G., 2012. A Heckman selection-t model. Journal of the American Statistical Association 107(497), 304–317.

[26] Marra, G., Wood, S.N., 2012. Coverage properties of confidence intervals for generalized additive model components. Scandinavian Journal of Statistics 39(1), 53–74.

[27] Marra, G., Radice, R., 2013a. Estimation of a regression spline sample selection model. Computational Statistics and Data Analysis 61, 158–173.

[28] Marra, G., Radice, R., 2013b. A Penalized Likelihood Estimation Approach to Semiparametric Sample Selection Binary Response Modelling. Electronic Journal of Statistics 7, 1432–1455.

[29] Marra, G., Radice, R., Wojtyś, M., Wyszynski, K., 2016. SemiParSampleSel: semiparametric sample selection modelling. R package version 1.3.

[30] McGovern, M.E., Barnighausen, T., Marra, G., Radice, R., 2015. On the Assumption of Joint Normality in Selection Models: A Copula Approach Applied to Estimating HIV Prevalence. Epidemiology, 26(2), 229–237.

[31] Mealli, F., Pacini, B., 2008. Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness. Computational Statistics and Data Analysis 53(2), 507–516.

[32] Miranda, A., Rabe-Hesketh, S., 2006. Maximum likelihood estimation of endogenous switching and sample selection models for binary, ordinal, and count variables. Stata Journal 6(3), 285–308.

[33] Nikoloulopoulos, A.K., Karlis, D., 2010. Regression in a copula model for bivariate count data. Journal of Applied Statistics 37(9), 1555–1568.

[34] Nocedal, J., Wright, S.J., 2006. Numerical Optimization. Springer-Verlag, New York.

[35] Omori, Y., Miyawaki, K., 2010. Tobit model with covariate dependent thresholds. Computational Statistics and Data Analysis 54(11), 2736–2752.

[36] Pigini, C., 2012. Of Butterflies and Caterpillars: Bivariate Normality in the Sample Selection Model. Università Politecnica delle Marche, Working paper No. 377.

[37] Puhani, P., 2000. The Heckman correction for sample selection and its critique. Journal of Economic Surveys 14(1), 53–68.

[38] R Development Core Team, 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[39] Radice, R., Marra, G., Wojtyś, M., in press. Copula regression spline models for binary outcomes. Statistics and Computing.

[40] Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press, London.

[41] Smith, M.D., 2003. Modelling sample selection using Archimedean copulas. The Econometrics Journal 6(1), 99–123.

[42] Smyth, G.K., Dunn, P.K., 1996. Randomized quantile residuals. Journal of Computational and Graphical Statistics 5(3), 236–244.

[43] Sklar, A., 1959. Fonctions de répartition é n dimensions et leurs marges. Publications de l'Institut de Statistique de l'Université de Paris 8, 229-231.

[44] Stasinopoulos, D.M., Rigby, R.A., 2005. Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C 54(3), 507–554.

[45] Stasinopoulos, D.M., Rigby, R.A., 2007. Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software 23(7), 1–46.

[46] O'Sullivan, F., Yandell, B.S., Raynor Jr., W.J., 1986. Automatic smoothing of regression functions in generalized linear models. Journal of the American Statistical Association 81(393), 96–103.

[47] Terza, J.V., 1998. Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. Journal of Econometrics 84(1), 129–154.

[48] Trivedi, P.K., Zimmer, D.M., 2007. Copula Modeling: An Introduction for Practitioners. Now Publishers Inc.

[49] United States Veterans Administration, 2001. National Survey of Veterans.

[50] Wiesenfarth, M., Kneib, T., 2010. Bayesian geoadditive sample selection models. Journal of the Royal Statistical Society Series C 59(3), 381–404.

[51] Wood, S.N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99(467), 673-686.

[52] Wood, S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman & Hall, London.

[53] Wood, S.N., 2013. On p-values for smooth components of an extended generalized additive model. Biometrika 100, 221–228.

[54] Wojtyś, M., Marra, G., Radice, R., in press. Copula Regression Spline Sample Selection Models: The R Package SemiParSampleSel. Journal of Statistical Software.

[55] Yee, T., Wild, C., 1996. Vector Generalized Additive Models. Journal of the Royal Statistical Society Series B 58(3), 481-493.

Appendix

1 Some numerical aspects

Starting values for the parameters of the selection equation are based on univariate probit model estimates. Starting values for the outcome equation parameters are instead obtained using a `gamlss`-like function implemented in `SemiParSampleSel` that allows to fit univariate models based on discrete distributions using selected observations only. The dependence between the selection and outcome equations is modeled through θ^* which is a monotonic transformation of θ. This is convenient since θ^* is not bounded by its parameter space and hence constrained optimization is not required. Table 9 shows θ^* defined in terms of θ for each copula. A starting value for θ^* is obtained using the two-stage

Copula	θ^*
FGM	$\tanh^{-1}(\theta)$
Normal	$\tanh^{-1}(\theta)$
AMH	$\tanh^{-1}(\theta)$
Clayton	$\log(\theta - \epsilon)$
Frank	$\theta - \epsilon$
Gumbel	$\log(\theta - 1)$
Joe	$\log(\theta - 1 - \epsilon)$

Table 9: Parameter θ^* defined in terms of θ. The values of θ^* are corrected with ϵ for Clayton, Frank and Joe to prevent the optimization algorithm from reaching boundary values of θ which are not included in the parameter space. ϵ is set to 10^{-8}.

Heckman-type approach discussed in Marra & Radice (2013a). This procedure yields an estimate of the correlation between the selection and outcome equations which is then transformed into θ^* depending on the copula employed.

In (2.4.3), when $y_2 = 0$ we use $F_2(y_2 - 1) = F_2(y_2) - f_2(y_2)$, case in which the value returned would be zero. This avoids evaluating $F_2(y_2 - 1)$ at a negative value. To carry out the optimization on $\mathbb{R}^p$ we also use the transformation $\sigma^* = \log(\sigma)$. When employing a Delaporte marginal ν is transformed using the logistic function. For the mean parameter we have that $\mu = \exp(\eta_2)$. In some cases, we have set precision bounds on the cdfs and copulae to avoid overflows. For the Poisson inverse Gaussian and Sichel distributions, the derivatives of the Bessel functions (see Table 2) are often difficult to evaluate. In these cases, numerical derivatives are used instead.

The two-step Inference Function for Margins (IFM) approach was explored in the initial phase of the project. Using this approach simplified the implementation considerably but did not unfortunately result in stable and efficient computations for the class of models considered in this paper. Under IFM certain key quantities needed in optimization could be calculated more rapidly, however more iterations were required to reach convergence due to the absence of cross-derivative information. Moreover, the algorithm would easily get stuck in local minima or maxima. The implementation proposed in this paper uses all derivative information and during testing showed to be superior to IFM in terms of stability and efficiency.

2 Derivation of smoothing parameter criterion

A sensible criterion to estimate smoothing parameters is (e.g., Wood, 2006)

$$E(\kappa) \;=\; E\left(\frac{1}{n^*}||\boldsymbol{\mu}-\tilde{\boldsymbol{\mu}}||^2\right), \tag{2.1}$$

where $n^* = 5n$ for a three parameter discrete distribution. Intuitively, by taking the difference between the true model predictions, $\boldsymbol{\mu}$, and prediction estimates, $\tilde{\boldsymbol{\mu}}$, we are minimizing the distance between the true and estimated models. However, $\boldsymbol{\mu}$ is unknown but it can be estimated. Let $\tilde{\boldsymbol{\mu}} = \sqrt{\mathbf{W}}\mathbf{X}\tilde{\boldsymbol{\delta}}$ and $\boldsymbol{\mu} = \sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\delta}$. Expanding κ in (2.1) yields

$$\begin{aligned}
\kappa &= \frac{1}{n^*}||\boldsymbol{\mu}-\tilde{\boldsymbol{\mu}}||^2 \\
&= \frac{1}{n^*}||\sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\delta}-\sqrt{\mathbf{W}}\mathbf{X}\tilde{\boldsymbol{\delta}})||^2 \\
&= \frac{1}{n^*}||\sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\delta}-\sqrt{\mathbf{W}}\mathbf{A}\mathbf{z}||^2 \\
&= \frac{1}{n^*}||\sqrt{\mathbf{W}}\mathbf{z}-\sqrt{\mathbf{W}}\mathbf{A}\mathbf{z}-\boldsymbol{\epsilon}||^2 \\
&= \frac{1}{n^*}\left(||\sqrt{\mathbf{W}}\mathbf{z}-\sqrt{\mathbf{W}}\mathbf{A}\mathbf{z}||^2 - 2\boldsymbol{\epsilon}^{\mathsf{T}}(\sqrt{\mathbf{W}}\mathbf{z}-\sqrt{\mathbf{W}}\mathbf{A}\mathbf{z}) + \boldsymbol{\epsilon}^{\mathsf{T}}\boldsymbol{\epsilon}\right) \\
&= \frac{1}{n^*}\big(||\sqrt{\mathbf{W}}\mathbf{z}-\sqrt{\mathbf{W}}\mathbf{A}\mathbf{z}||^2 - 2\boldsymbol{\epsilon}^{\mathsf{T}}\boldsymbol{\epsilon} - 2\boldsymbol{\epsilon}^{\mathsf{T}}\mathbf{W}\mathbf{X}\boldsymbol{\delta} + 2\boldsymbol{\epsilon}^{\mathsf{T}}\sqrt{\mathbf{W}}\mathbf{A}\sqrt{\mathbf{W}}^{-1}\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^{\mathsf{T}}\sqrt{\mathbf{W}}\mathbf{A}\mathbf{X}\boldsymbol{\delta} \\
&\quad +\boldsymbol{\epsilon}^{\mathsf{T}}\boldsymbol{\epsilon}\big) \\
&= \frac{1}{n^*}\left(||\sqrt{\mathbf{W}}\mathbf{z}-\sqrt{\mathbf{W}}\mathbf{A}\mathbf{z}||^2 - \boldsymbol{\epsilon}^{\mathsf{T}}\boldsymbol{\epsilon} - 2\boldsymbol{\epsilon}^{\mathsf{T}}\mathbf{W}\mathbf{X}\boldsymbol{\delta} + 2\boldsymbol{\epsilon}^{\mathsf{T}}\sqrt{\mathbf{W}}\mathbf{A}\sqrt{\mathbf{W}}^{-1}\boldsymbol{\epsilon} + 2\boldsymbol{\epsilon}^{\mathsf{T}}\sqrt{\mathbf{W}}\mathbf{A}\mathbf{X}\boldsymbol{\delta}\right),
\end{aligned}$$

where $\boldsymbol{\epsilon} = \sqrt{\mathbf{W}}\mathbf{z} - \sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\delta}$ and $\sqrt{\mathbf{W}}\mathbf{A} = \sqrt{\mathbf{W}}\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X}+\mathbf{S}_\lambda\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}$. Note that

$$\sqrt{\mathbf{W}}\mathbf{z} \;=\; \sqrt{\mathbf{W}}\left(\mathbf{W}^{-1}\mathbf{u}+\mathbf{X}\boldsymbol{\delta}\right).$$

Since $E(\mathbf{u}) = \mathbf{0}$,

$$\begin{aligned}
E\left(\sqrt{\mathbf{W}}\mathbf{z}\right) &= E\left(\sqrt{\mathbf{W}}\left(\mathbf{W}^{-1}\mathbf{u}+\mathbf{X}\boldsymbol{\delta}\right)\right) \\
&= \sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\delta}.
\end{aligned}$$

Also, we have $\mathrm{Var}(\mathbf{u}) = E(\mathbf{u}\mathbf{u}^{\mathsf{T}}) - (E(\mathbf{u}))^2 = E(\mathbf{u}\mathbf{u}^{\mathsf{T}}) = \mathbf{W}$. Hence,

$$\begin{aligned}
\mathrm{Var}\left(\sqrt{\mathbf{W}}\mathbf{z}\right) &= \mathrm{Var}\left(\sqrt{\mathbf{W}}\left(\mathbf{W}^{-1}\mathbf{u}+\mathbf{X}\boldsymbol{\delta}^*\right)\right) \\
&= \mathrm{Var}\left(\sqrt{\mathbf{W}}^{-1}\mathbf{u}+\sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\delta}^*\right) \\
&= \mathrm{Var}\left(\sqrt{\mathbf{W}}^{-1}\mathbf{u}\right) \\
&= \sqrt{\mathbf{W}}^{-1}\mathbf{W}\sqrt{\mathbf{W}}^{-1} \\
&= \mathbf{I}.
\end{aligned}$$

From likelihood theory we also know that the distribution of the score vector is normal. Therefore,

$$\begin{aligned}
\sqrt{\mathbf{W}}^{-1}\mathbf{u} &\sim N(\mathbf{0},\mathbf{I}),\\
\mathbf{W}^{-1}\mathbf{u} &\sim N(\mathbf{0},\mathbf{W}^{-1}),\\
\mathbf{W}^{-1}\mathbf{u}+\mathbf{X}\boldsymbol{\delta} &\sim N(\mathbf{X}\boldsymbol{\delta},\mathbf{W}^{-1}),\\
\mathbf{z} &\sim N(\mathbf{X}\boldsymbol{\delta},\mathbf{W}^{-1}),\\
\sqrt{\mathbf{W}}\mathbf{z} &\sim N(\sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\delta},\mathbf{I}),
\end{aligned}$$

or

$$\sqrt{\mathbf{W}}\mathbf{z}-\sqrt{\mathbf{W}}\mathbf{X}\boldsymbol{\delta} \sim N(\mathbf{0},\mathbf{I}).$$

Taking the expectation of κ yields

$$\begin{aligned}
E(\kappa) &= \frac{1}{n^*}E\left(||\sqrt{\mathbf{W}}\mathbf{z}-\sqrt{\mathbf{W}}\mathbf{A}\mathbf{z}||^2\right)-\frac{1}{n^*}E\left(\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon}\right)-\frac{1}{n^*}E\left(2\boldsymbol{\epsilon}^\mathsf{T}\mathbf{W}\mathbf{X}\boldsymbol{\delta}\right)\\
&+ \frac{1}{n^*}E\left(2\boldsymbol{\epsilon}^\mathsf{T}\sqrt{\mathbf{W}}\mathbf{A}\sqrt{\mathbf{W}}^{-1}\boldsymbol{\epsilon}\right)+\frac{1}{n^*}E\left(2\boldsymbol{\epsilon}^\mathsf{T}\sqrt{\mathbf{W}}\mathbf{A}\mathbf{X}\boldsymbol{\delta}\right).
\end{aligned}$$

The third and the fifth term are equal to zero since $E(\boldsymbol{\epsilon})=\mathbf{0}$. Also,

$$E(\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon}) = E\left(\sum_{i=1}^{n^*}\epsilon_i^2\right)=n^*$$

and

$$\begin{aligned}
E(\boldsymbol{\epsilon}^\mathsf{T}\sqrt{\mathbf{W}}\mathbf{A}\sqrt{\mathbf{W}}^{-1}\boldsymbol{\epsilon}) &= E(\mathrm{tr}\left(\boldsymbol{\epsilon}^\mathsf{T}\sqrt{\mathbf{W}}\mathbf{A}\sqrt{\mathbf{W}}^{-1}\boldsymbol{\epsilon}\right))\\
&= E(\mathrm{tr}\left(\sqrt{\mathbf{W}}\sqrt{\mathbf{W}}^{-1}\mathbf{A}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\mathsf{T}\right))\\
&= \mathrm{tr}\left(\mathbf{A}E\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\mathsf{T}\right)\right)) \qquad (2.2)\\
&= \mathrm{tr}\left(\mathbf{A}\mathbf{I}\right)\\
&= \mathrm{tr}\left(\mathbf{A}\right).
\end{aligned}$$

The first line of (2.2) is justified by the fact that a scalar is its own trace. The remaining lines follow from the properties of the trace of a matrix. Thus,

$$E(\kappa) = \frac{1}{n^*}E\left(||\sqrt{\mathbf{W}}\mathbf{z}-\sqrt{\mathbf{W}}\mathbf{A}\mathbf{z}||^2\right)-1+\frac{2}{n^*}\,\mathrm{tr}(\mathbf{A})$$

This can be estimated as (Wood, 2006)

$$\mathcal{V}_u(\boldsymbol{\lambda}) = \frac{1}{n^*}||\sqrt{\mathbf{W}}(\mathbf{z}-\mathbf{X}\tilde{\boldsymbol{\delta}})||^2-1+\frac{2}{n^*}\,\mathrm{tr}(\mathbf{A}),$$

where $n^* = 5n$, $\sqrt{\mathbf{W}}\mathbf{A} = \sqrt{\mathbf{W}}\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \mathbf{S}_\lambda\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}$ is the hat matrix.

3 Software implementation

The software for implementing all the model features, estimation and inferential procedures outlined in this paper is freely available through the `R` package `SemiParSampleSel` (Marra et al., 2016). The framework this paper provides can allow researchers and policymakers to apply a transparent approach to account for systematic non participation in their data. The features of this software have been designed specifically with transparent and straightforward dissemination of results in mind. The main function is

```
SemiParSampleSel(list(y.sel ~ s(x1) + x2 + x3, y ~ s(x1) + x2, ~ x3),
                 data = dataset, BivD = "N", margins = c("probit", "P"), ...)
```

The first argument is a list of at least two formulae. In this case, `x1`, `x2` and `x3` denote the covariates, where the effects of `x1` are modeled using thin plate regression splines. `y.sel` and `y` are the binary selection and observed discrete outcome variables which are modeled using normal and Poisson distributions (`margins = c("probit", "P")`). The copula model is set to Gaussian (`BivD = "N"`). The list of formulae can be augmented by an extra equation for θ which allows the user to specify a predictor for the copula dependence parameter. More equations can be included for modeling σ and ν as functions of covariates. The choice of outcome distribution can be changed, for instance, to negative binomial (`margins = c("probit", "NB")`). A Frank copula is employed when `BivD = "F"`.

Post-estimation, QQ-plots can be obtained using `post.check()` which are based on the sample selection model estimates. Prior to fitting `resp.check()` can help decide roughly which distributions are likely to produce the best post-fitting results, however it is worth stressing that the resulting plots would be based on univariate models (ignoring selectivity) and hence may lead to misleading conclusions. Model `summary` and `plot` functions work in a similar fashion as those of generalized linear and additive models. Predicted averages, with corresponding intervals, can be obtained using the `aver` function. Convergence can be checked using `conv.check()`. More details and options can be found in the documentation of `SemiParSampleSel`.